CS TRACK
Investigating Citizen Science

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

Web Analytics – Final Report
D3.3

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

Title of project	CS Track
Full title of project	Expanding our knowledge on Citizen Science through analytics and analysis
Title of this document	Web Analytics – Final Report
Number of this document	D3.3
Dissemination level	Public
Due date	30th November 2022
Actual delivery	
Versioning history	Version 1.0
Authors	Core authors: RIAS - Simon Krukowski (editor), H. Ulrich Hoppe, Cleo Schulten URJC - Fernando Martínez Martínez UPF - Miriam Calvera Isábal, Nicolás Gutiérrez Páez Additional contributors: Aaron Peltoniemi (JYV - internal review), Nils Malzahn (RIAS), Patricia Santos Rodríguez (UPF), David Roldán Álvarez (URJC), Estefanía Martín Barroso (URJC)
Executive summary	This deliverable wraps up the work on “web-based analytics” in WP3 in addition to the previous deliverables on “methods and tools” (D3.1) and on design and implementation of the Analytics Workbench and tools (D3.2). A first add-on is the usage and evaluation experience with the AWB, which is based on several interactive presentations, workshops and webinars. These, in turn, served as a source of data to evaluate this tool suite from a user interface, application and algorithmic perspective. Several case studies included here represent the previously introduced micro, meso and macro perspectives (see D3.1 and D1.2). In spite of the difference in scope, the macro perspective (Twitter analyses) and the micro perspective on participation and engagement share the use of network models and network analysis techniques, whereas the meso level analyses are content-centric. Regarding the interfacing and exportation of analytics results, we address several channels including the information export through the eMagazine as well as through project workshops / webinars and “normal” scientific publications. For the technical outcomes, we report on our open source and open data strategy that relies on open archives and code repositories.

Table of Contents

List of Abbreviations

Abbreviation	Definition
API	*Application programming interface*
AWB	*Analysis workbench*
BERT	*Bidirectional encoder representations from transformers*
CS	*Citizen science*
DK	*Denmark*
ESA	*Explicit semantic analysis*
LDA	*Latent Dirichlet allocation*
ML	*Machine learning*
MOOC	*Massive open online course(s)*
NCD	*Noncommunicable diseases*
NER	*Named entity recognition*
NESA	*Non-orthogonal ESA*
RA	*Research area*
RT	*Retweet*
SDG	*Sustainable development goal*
SNA	*Social network analysis*
SQL	*Structured query language*
STEM	*Science, technology, engineering, and mathematics*
TF-IDF	*Term frequency - inverse document frequency*
UEQ	*User experience questionnaire*
UMAP	*Uniform manifold approximation and projection*
WP	*Work package*
XESA	*EXtended Explicit Semantic Analysis*

Section 1: Orientation and overview

The methodological and theoretical basis of web analytics in CS Track has been laid out in D3.1. The approaches adopted in the project fall into the two main categories: content-based analyses and network-based analyses (including network modelling techniques). A second organising principle is the distinction of three levels of analysis: *micro*, *meso* and *macro*. This distinction had already been introduced in D3.1 and an updated specification has been included in D1.2 (see section 6.3). The following Table 1 describes the levels and characterises typical applications:

Table 1: Characteristics of the three analysis levels used in CS Track

Level	Scope	Tools / Methods	Example
Micro	Based on small samples using specific selection criteria or hand-picking	Discourse analysis or social network analysis (SNA) of project-specific web resources; may involve manual coding of data and a mixture of quantitative and qualitative methods	Participation and distribution of roles, COVID CS projects (see section 4.1)
Meso	Homogeneous collections of projects selected from the WP2 database by their metadata (or at random)	Information extraction from homogeneously formatted datasets - linguistic analyses (e.g., NER, ESA, LDA) - descriptive statistics	Identification of research areas or SDGs based on project descriptions, assessment of multi-disciplinarity
Macro	Open collections of projects or initiatives, not necessarily already contained in the CS Track database	Harvesting and analysis of social media data (e.g., Twitter) - (social) network analysis (e.g., centrality dynamics in retweet networks) - linguistic analyses (e.g., keyword extraction)	Calculation of public outreach, “popularity”, “prominence”, trending topics in CS

This deliverable document (D3.3) wraps up the work on “web-based analytics” in WP3. Whereas the previous deliverables dealt with “methods and tools” in a conceptual, forward-looking perspective (D3.1) and with the basic analytics toolset and the Analytics Workbench (AWB) from a system design and implementation perspective (D3.2), D3.3 adds the usage and evaluation experience with the AWB in the line of Task 3.2 and covers the work on “case studies” (Task 3.3) and “interfacing and exportation of results” (Task 3.4).

The AWB has been used as an interactive demonstrator in several interactive presentations, workshops and webinars, which in turn served as a source of data to evaluate the qualities and possible deficits of this tool from a user interface / user experience and application perspective.

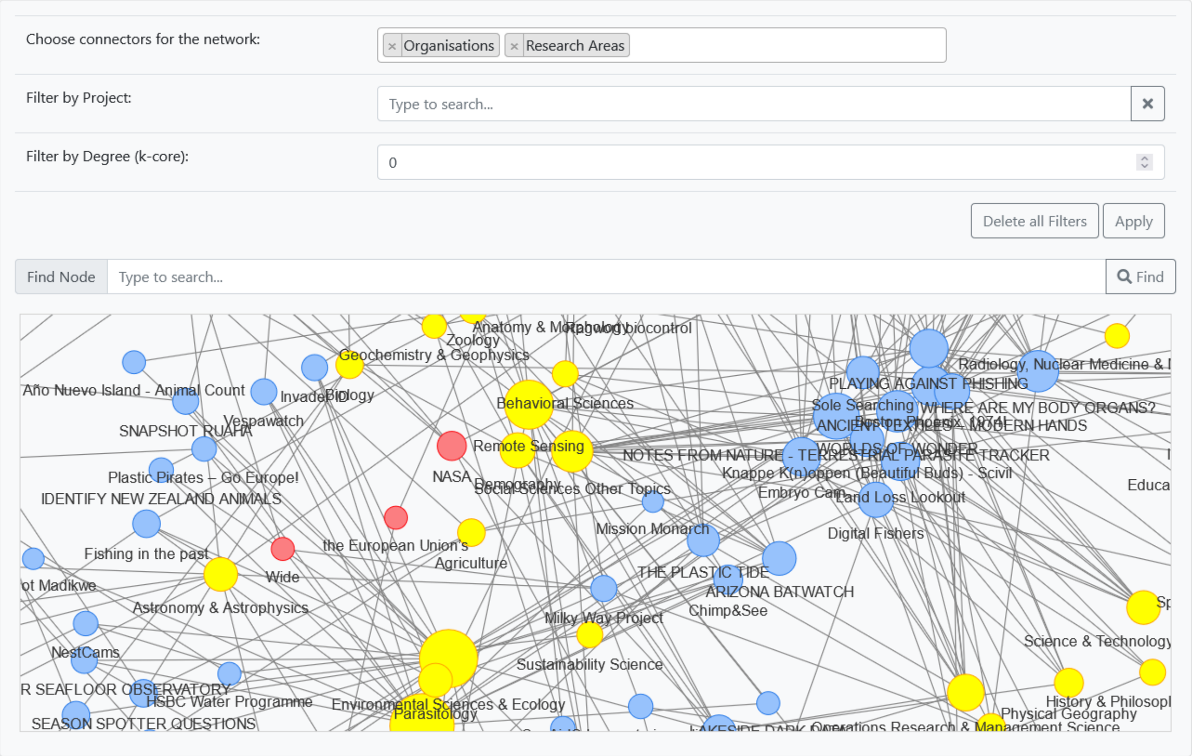
Together with the AWB we have also exposed and tested the information extraction algorithms (NER and ESA) based on collections of project descriptions. The workshop feedback allowed us to improve on these algorithms. These experiences and efforts have been documented in section 2.

The scope of the case studies covered in section 3 corresponds to the map laid out in Table 1 (see above). The three subsections represent the meso, micro and macro perspectives (in this order). In spite of the difference in scope, the macro perspective (Twitter analyses) and the micro perspective on participation and engagement share the use of network models and network analysis techniques, whereas the meso level analyses are content-centric. Notably, the analyses of participation and engagement that are based on traces of forum interactions have reached a level of standardization and automation that makes them quite easily transferable.

The interfacing and exportation of analytics results (T3.4) had three main directions: (1) The feeding back of results to the database of projects originally developed and built up in WP2, (2) the usage of tools with example data in project workshops and webinars, and (3) the publication of results in the eMagazine, especially in the form of graphical articles allowing for data visualisation. This is taken up in section 4, which also includes our strategies for data sharing (open source and open data) and publications.

Section 2: Analytics workbench

In the line of content-analytic approaches, we have focused on “Explicit Semantic Analysis” (ESA - Gabrilovich & Markovitch, 2007) as a technique to capture semantic features related to textual documents such as project descriptions. This is particularly used to compute the association of research areas (RAs) or sustainable development goals (SDGs) to given projects represented by their textual descriptions captured in the CS Track database. This functionality is included with the “Analytics Workbench” (AWB) that has been described in detail in D3.2. The AWB provides interactive access to different content analysis techniques. Beyond the ESA-based calculation of RA and SDG associations it also allows for extracting named entities (person names, geographical entities, organisations, etc.). “Named Entity Recognition” (NER) is particularly relevant for the standardisation of terminology and for anonymisation. In addition to giving access to content analysis functions, the AWB allows for navigation of the database based on semantic similarities or bridges and provides network visualisations. Figure 1 illustrates this using a sample of projects.



Choose connectors for the network: × Organisations × Research Areas

Filter by Project: Type to search...

Filter by Degree (k-core): 0

Figure 1: AWB dashboard with a network view over the collected data in the lower half (blue - projects; red - organisations; yellow - research areas)

In the rest of this section, we describe general experiences and extensions of methods related to ESA and to the AWB. Reports on actual applications to specific research questions are part of section 3.

2.1. Validation and adaptation of ESA

The technique of ESA has already been introduced in D3.1. ESA derives semantic relatedness based on a precalculated inverted index built by using an encyclopaedia such as Wikipedia. Concepts to be detected are represented by their corresponding Wikipedia pages (text), i.e., the encyclopaedia serves as a reference knowledge base or ontology. The relatedness or association strength of any other given text to a concept is calculated by text similarity. However, the actual similarity calculation does not use the text sources of the Wikipedia reference pages but pre-processed versions in the form of attribute vectors that capture occurring terms with term weights that correspond to standard TF-IDF values. TF-IDF stands for term frequency–inverse document frequency, this is a numerical statistic calculated to reflect how important a word is in a document which is inside a collection of documents. Given these vectors, an inverted index is created, mapping terms to the concepts in which they appear. In this context, these terms are usually reduced to their stem form. In setting up the inverted index it is also possible to filter out term-concept relations of low significance. The term vectors from the inverted index can be used to determine the relatedness of two terms by calculating the cosine similarity of the vectors. This translates to the idea that terms that co-occur in multiple articles are more similar to each other than terms that do not co-occur in articles.

This procedure can then be upscaled to entire texts by first summing up the term vectors of all occurring words to a text vector for each text and then using those to calculate the cosine similarity. This allows for matching other texts against the selected concepts represented by their Wikipedia pages. Here, the other text serves as a kind of query posed against the pre-established body of concepts. To improve performance and reduce calculation time, the words appearing in the query text can be filtered first. While it is common practice to discard stopwords in this context, it is also possible to filter them by TF-IDF scores. In this case, only the stems of the remaining words are then used to determine word vectors, which, in turn, can be used to calculate the text vector.

In recent years, various approaches to improving ESA have been implemented and reported. In their paper on eXtended Explicit Semantic Analysis (XESA), (Scholl et al., 2010) report on their approaches to enrich ESA using Wikipedia’s link graph, a category structure, or a combination of both. In their evaluation of the results using a methodology to evaluate search engine rankings, they conclude that the mixed approach leads to worse results than pure ESA. Conversely, the other approaches improve ESA, with the article graph variant bringing the best results (Scholl et al., 2010). Thematically Reinforced ESA, introduced by Haralambous and Klyuev (2014), uses an extension for the TF-IDF values used in the ESA matrix. Adding to the standard TF/IDF calculation, they extend it by taking into account whether or not the term occurs in the ancestors of each given page, indicating an additional value of the term in a category. Using a text classification task as a benchmark, they found that their approach performed significantly better than the standard ESA (Haralambous & Klyuev, 2014). The Non-Orthogonal ESA (NESA) approach by Aggarwal et al. (2015) aimed to introduce the possibility that two terms are similar to each other without sharing many co-occurrences – like football and soccer – into ESA by including concept relatedness measures into the ESA approach. For their evaluation, they use six benchmarks for word relatedness and find that each of their four NESA approaches outperforms the standard ESA method (Aggarwal et al., 2015).

Our ESA calculations use a pre-compiled inverted index from a Wikipedia dump as of April 2019 that is stored in an SQLite database. We used Web of Science’s research area classification of 152 research areas[1] as a basis for our taxonomy and thus for the selection of Wikipedia reference pages matching the items in the taxonomy. While this was usually based on one-to-one matches, we had to use multiple Wikipedia articles for some. One such example is the research area “Film, Radio & Television” which was matched with the articles for “Film”, “Radio” and “Television”. This table was then used to calculate the corresponding text vectors and store them in an additional database.

As previously described, it is advisable to filter the texts, which, in this case, applies to research areas as well as project descriptions. Additionally, to the standard elimination of stopwords, we decided to employ TF-IDF filtering. For this, we calculate the TF-IDF scores for a text’s words and determine the

[1] https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html

highest score in the text. We then use 20% of the highest score as the minimum score with which a word still becomes included to calculate the text vector.

The 20% mark was chosen based on trial and error with different values. For this, we used 10 projects for which we assigned research areas manually. Then we used ESA to assign research areas to these projects in variants with different thresholds (10%, 20%, 25%, 30%, 40% and 50%). The calculation was done with precalculated vectors for the research areas where only stopwords were excluded. A preliminary overview of the results in the light quality measures (precision, recall, F1 score) narrowed the window of plausible choices to 20-25% as a cut-off. In the next step, we included another 10 projects with additional manual assignments, yielding the following scores for precision, recall and F1 score:

	Precision	Recall	F1 Score
TF-IDF-Cut-off at 20%	0.545	0.214	0.308
TF-IDF-Cut-off at 25%	0.455	0.179	0.256

An additional cut-off had to be set to determine which research areas are assigned to a project based on their calculated similarities. Since the similarity of texts is rarely zero, such a cut-off is imperative. To account for the inter- and multi-disciplinarity of projects, it is not possible to use the best *N* research areas and the range of different highest similarities does not allow for a fixed cut-off. Thus, we chose to use a cut-off relative to the highest reached similarity per project. This ensures that at least one research area and possibly more are assigned. This cut-off was set to 75% based on trials.

After these initial decisions, we iteratively improved our approach. From the start, we included TF-IDF thresholds to focus on the most meaningful text elements. Building on this, we chose to improve our usage of the standard ESA approach. For this purpose, we enriched the ESA-based assignments with assignments based on DBPedia matches for the Wikipedia pages of the research areas in the project descriptions. Additionally, we reworked the set-up of the precalculated research area vectors, to save the current Wikipedia pages that are used and enable us to edit those where needed – i.e., removing noise – before the vectors are calculated. Lastly, we enriched the original list of research areas by joining it with another classification by Web of Science which has a total 252 research areas[2] – bringing us to a list of 163 research areas. Based on these added research areas we corrected our original manual research area assignment and recalculated precision, recall and F1 score for the new set-up.

	Precision	Recall	F1 Score
New set-up with 20% TF-IDF cut-off	0.480	0.313	0.379

[2] https://incites.help.clarivate.com/Content/Research-Areas/wos-research-areas.htm

2.2. Evolution and usage of the AWB

To evaluate, test and disseminate the AWB, CS Track conducted two workshops targeting other researchers, institutions, and project managers. The second workshop figured as the first instance in the series of CS Track Webinars. Due to limitations induced by the prepared task materials both workshops had 30 registered participants from 14 (first) and 15 (second) countries. Both workshops started with a short presentation of the AWB including a demo before transitioning to a hands-on activity. In the first workshop, we prepared a structured online questionnaire that guided the participants through the usage of the workbench, instructing them to each analyse a predetermined project. For the second workshop, we switched to a more group-oriented approach and used breakout sessions where we prepared lists of related projects for each session for the participants to analyse. At the end of both workshops, we asked the participants to answer the short User Experience Questionnaire (UEQ-S; Schrepp et al., 2017) and a few questions about the helpfulness of individual workbench functionalities.

Furthermore, we ran an internal workshop to investigate to what extent the automatically generated results are deemed as correct or required manual revision by expert workbench users.

All of these workshops were run before the last iteration of improvements for the AWB took place.

2.2.1. Webinar feedback

Collectively, from both webinar workshops, we received replies for the UEQ-S from 18 participants, the results of which are shown in Figure 2.

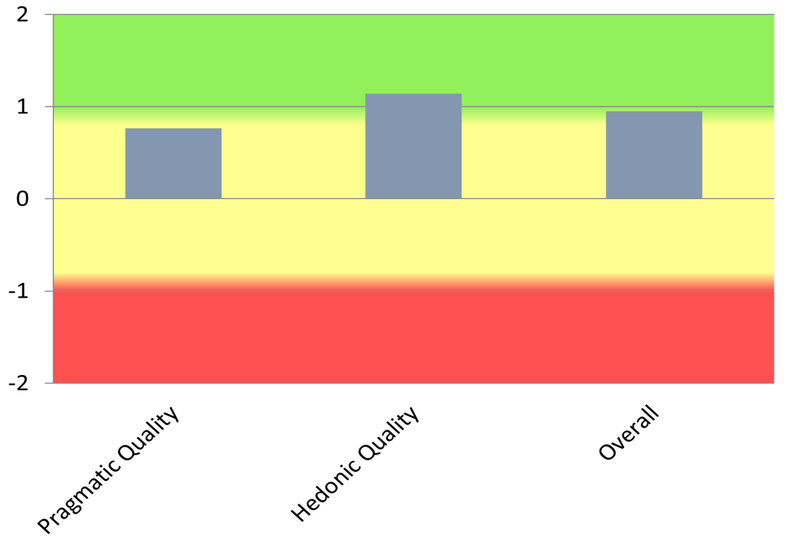


Figure 2: UEQ-S results

On average, the workbench received a score of 0.76 for pragmatic quality, a score of 1.14 for hedonic quality and a score of 0.95 for overall. The UEQ-S measures on a scale from -3 to 3, with values above 0.80 constituting a positive result. Based on some of the feedback in the internal workshop, we surmised that one reason for the low pragmatic score may have been due to the unintended complexity of the modification of ESA-based results to be unintuitive. This was therefore changed in the meantime. Regarding helpfulness, we asked the participants from the first webinar workshop (total of 13 participants) to rate each functionality according to its perceived helpfulness on a four-point scale: 0 – not helpful; 1 – somewhat helpful, 2 – helpful and 3 – very helpful. The results of this are displayed in Figure 3.

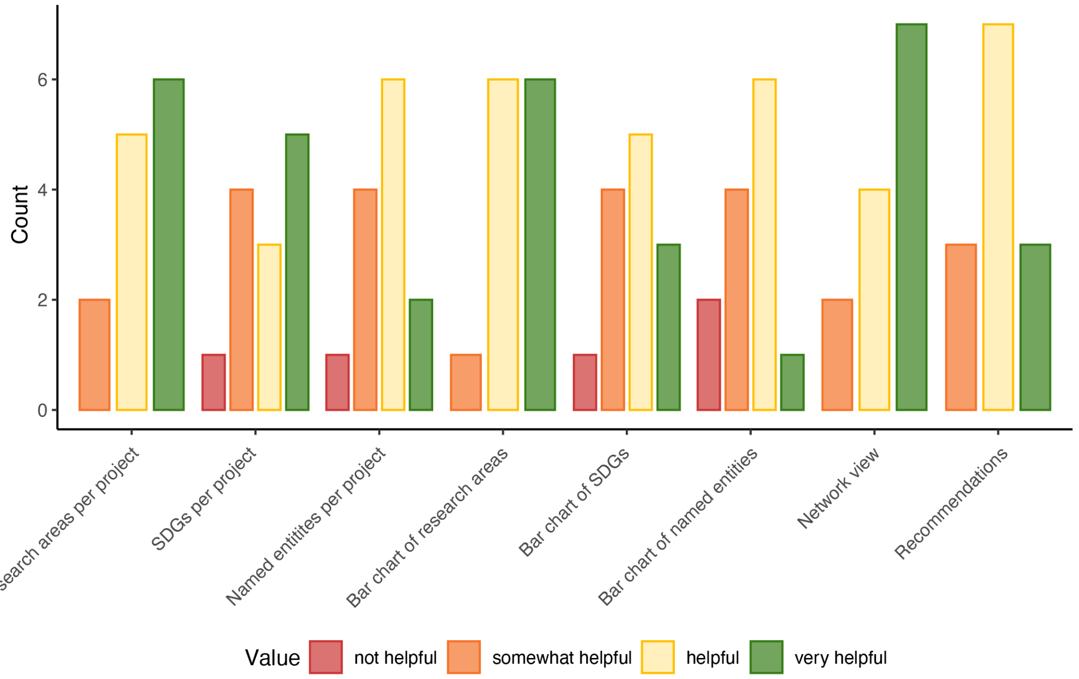


Figure 3: Helpfulness rating per functionality

As for the perceived helpfulness of project specific results, we asked separately for research areas, SDGs and named entities. Research area assignments were rated as somewhat helpful by two participants, as helpful by five, and very helpful by six. This translates to a mean of *M* = 2.31 (*SD* = 0.75), which can be interpreted as helpful slightly leaning towards very helpful. The SDGs rating reaches a mean of *M* = 1.92 (*SD* = 1.04) and the named entities have a mean of *M* = 1.69 (*SD* = 0.85). The research area bar chart received a mean of *M* = 2.38 (*SD* = 0.65) with one somewhat helpful rating, and six helpful and six very helpful. The SDG bar chart ranks lower with a mean of *M* = 1.77 (*SD* = 0.93). The named entity bar chart has a mean rating of *M* = 1.46 (*SD* = 0.88). The network view has two ratings of somewhat helpful, four of helpful and seven of very helpful which results in a mean of *M* = 2.38 (*SD* = 0.77). Lastly, the recommendations were classified as helpful by seven, somewhat helpful and similarly very helpful by three each, resulting in an average of *M* = 2.00 (*SD* = 0.71).

Although the tasks the participants were asked to complete covered research areas, SDGs and named entities in a comparable way with equivalent visual representations, we received quite different ratings of the perceived helpfulness. It is plausible to assume that the differences are not due to usability features but have to do with the subjective feeling of relevance related to these items in the task context. The least appreciated feature are the named entities with a mean of *M = 1.69* (*1.46* for the bar chart given a ranked overview). Indeed, there is "noise" in these results, including, e.g., unspecific number words (like "thousands"). The maximum value of *M = 2.31* (*2.38* for the overview bar chart) for research areas indicates that these are seen as most relevant for characterising the projects. With *M = 1.92* (*1.77* for the bar chart view), SDGs fall in between. They are likely seen as less specific or characteristic for classifying projects. The network view that facilitates navigation in the space of projects is received very positively (*M = 2.38*), whereas the perceived helpfulness of project recommendations is less pronounced although still quite positive (*M = 2.0*).

2.2.2. User validation of the ESA-based association of RAs

Following the second workshop, we wanted to examine further to which extent users would modify the ESA-based results to fit their idea of the correct results. Since the projects we selected for the workshop were not representative of the projects in the database, as they were selected to be similar enough to find common denominators within each group, we chose to set up a separate list of projects. We used the CS Track project meeting in Barcelona (March 29-31, 2022) to run this experiment. For this, we prepared a list of 50 projects. We then ran two rounds in which the projects were analysed and the results verified, with the idea being that for each project we would then have 2 evaluators or raters, whose results we could then use to calculate a mean value. Due to it not being two raters rating all projects but rather 16 raters, it was not possible to calculate inter-rater reliability. Also, because some projects had duplicates in the database or related projects that were similarly named, we ended up with a list of 56 projects in our dataset, with 36 of them having been rated by 2 raters.

We then used this dataset to investigate each research area and SDG, how often they were assigned in the ESA-based results, which were removed or added manually. If the two raters for one project agreed on assigning a certain research area or SDG this was counted as 1. If they coincided in not assigning it, this was counted as 0. If only one rater decided to assign the item, this was counted as we counted as 0.5. Overall, 72 different research areas occurred in the dataset, i.e., they were assigned to at least one project. There were 192 ESA-based assignments, 84 manual removals and 58 manual additions resulting in 166 final assignments. This means that 56.25% of the ESA-based assignments were considered as correct and kept, and 65.06% of the final assignments were ESA-based. The top 10 ESA-based research areas can be seen in Table 2.

Table 2: Human confirmation/revision of RA assignments (top 10 RA by occurrence)

Research Area	Number of project assignments			
	ESA-based	Removed manually	Added manually	Final assignment
Remote Sensing	23	13	1	11
Biodiversity and Conservation	19	0	4.5	23.5
Environmental Sciences and Ecology	13	1.5	4.5	16
Parasitology	11	9	0.5	2.5
Astronomy and Astrophysics	9	1	4.5	12.5
Operations Research and Management Science	8	6	0	2
Development Studies	7	5.5	1	2.5
Radiology, Nuclear Medicine and Medical Imaging	7	5	0	2
Imaging Science and Photographic Technology	6	3.5	1.5	4
Geochemistry and Geophysics	6	2	0	4

11 research areas were assigned to more than 2 projects but were removed from more than half of them. The peak in terms of having the most ESA-based assignments manually removed was *Parasitology,* which was assigned 11 times and removed 9 times. *Remote Sensing* was automatically assigned the most (23 times) and manually removed the most as well (13 times). For two research areas (*Zoology* and *Computer Science*), we found that they had been automatically assigned at least

once and then manually assigned more times than automatically. 18 research areas had not been assigned ESA-based but were assigned manually 0.5 to 5.5 times. The three research areas that were assigned the most after manual changes (*Biodiversity and Conservation*, *Environmental Science and Ecology* and *Astronomy and Astrophysics*) were manually removed 0 to 1.5 times and manually added 4.5 times each.

All 17 SDGs occurred in the dataset, with 77 ESA-based assignments, 47.5 manual removals and 21 manual additions which resulted in 50.5 final assignments. So, for the SDGs 38.31% of ESA-assignments were kept and 58.42% of the final results are ESA-based. For seven SDGs, we found that they had been assigned to more than two projects but had more than half their assignments manually removed. The most notable was SDG #1, which was assigned 14 times based on ESA and removed manually 13 times. The most frequent SDG by far was SDG #15, both in ESA-based assignments and manually additions, and consequently also in final assignments. SDGs #14 and #10 were assigned once each ESA-based then each removed 0.5 times and manually added 4.5 and 1.5 times respectively, making these two the only SDGs that were assigned manually more often than ESA-based with at least 1 ESA-based assignment. Additionally, two SDGs – #2 and #6 – were not assigned ESA-based but were added manually 1 and 0.5 times respectively.

2.2.3. Comparison of user assignments with updated ESA assignments

A parallel thread of work related to ESA was a reworking the prepared textual references for RAs. This included additional RAs that were not part of original WoS taxonomy (such as *Ornithology* and *Political Science*) as well as a modification of some of the reference pages that had been identified to contain out-of-focus content. The corresponding vectors were recalculated. The manual assignments reported above were still based on the old ESA assignments. Having the new assignments, we are now able to compare the old ESA results with human- and new ESA-based assignments. The following Table 3 aggregates these results:

Table 3: Top 10 ESA-based assignments (old) and corresponding revisions and changes

Research Area	Number of project assignments				
	ESA-based (old)	Removed in new ESA	Added in new ESA	ESA-based (new)	Overlap of automatic and manual removals
Remote Sensing	23	4	0	19	2
Biodiversity and Conservation	19	3	0	16	0
Environmental Sciences and Ecology	13	3	2	12	1
Parasitology	11	6	0	5	5
Astronomy and Astrophysics	9	0	0	9	0
Operations Research and Management Science	8	3	1	6	3

Development Studies	7	4	0	3	4
Radiology, Nuclear Medicine and Medical Imaging	7	0	0	7	0
Geochemistry and Geophysics	6	1	0	5	0
Imaging Science and Photographic Technology	6	0	1	7	0

In the new dataset, we found 70 occurring research areas, of which 27 did not change in their assignments with the new assignment. Overall, we found 147 common assignments between the old and the new ESA-based results, 45 assignments were made by the old one but not the new one, and 51 vice versa.

16 entire research areas had not been assigned in the previous ESA-based assignment and were ‘added’ with the new one, which accounts for 38 new ESA-based assignments. Three of these newly occurring research areas were research areas that were newly added to the taxonomy with 13 total assignments. The most notable of these additions was *Ornithology* which was added 10 times.

The 45 assignments that were not supported by the new approach span over 21 research areas, which for 7 research areas meant that they were not represented in the new ESA-based assignments. The three research areas that were removed most in the new approach were *Parasitology* (6x), *Development Studies* (4x), and *Remote Sensing* (4x). A total of 16 research areas were only removed from projects in the switch and not added to different projects.

The changes in assignments can be attributed to three main factors: 1) the newly added research areas not only had assignments that were not possible previously, but they also influenced the similarity ranking and potentially the cut-off which may lead to other research areas not being assigned; 2) we modified the reference text for *Remote Sensing* because we noticed an entire paragraph being off-topic, this may have led to fewer faulty assignments of *Remote Sensing* which in turn may have made other – more favourable – assignments possible; and 3) in our modifications, we detected some HTML artefacts that were left in the reference texts after they were pulled from Wikipedia, so we had modified our algorithm to remove those before the text vectors were calculated. It is possible, that these artefacts influenced the results, and thus their removal led to slightly modified assignments.

Section 3: Case studies

According to the description of task T3.3, case studies should foster the integration and evaluation of developed tools and help to test and characterise the results that can be achieved by applying computational techniques to publicly available sources. These findings complement results gained on the basis of questionnaires and interviews. As such, they are also an important source for generating policy recommendations. In addition to addressing relevant topics of interest, the case studies have been designed to cover the different levels of analysis and the main methodological approaches already introduced in D3.1.

The following sections will first report on results based on content analysis through ESA and the "Analytics Workbench" (AWB). These shed light on the (inter-)disciplinary nature of CS projects as well as on the interrelation of disciplinary orientation with the relevance to certain SDGs. Aspects of participation and motivation in CS projects are investigated by analysing detailed data from project forum and talk pages using SNA techniques. Although this does not allow for directly assessing motivation as such, we can gain insights into the interplay of activities and incentives especially on the part of volunteers. According to the distinction of levels of analysis in Table 1 (Section 1), the *macro* level analyses were based on a large corpus of Twitter data, which was analysed using topic analysis, machine learning and again SNA approaches. This has revealed characteristics of the discourse in different fields of intersection of CS with certain applications domains such as eHealth, education, and sustainable development.

3.1. SDGs and research areas

This first section is dedicated to applications of ESA and the workbench in the content-analytic line of work, which belongs to the *meso* level of analysis. The specific attributes are research areas (RAs) and sustainable development goals (SDGs) susceptible to the usage of ESA based on their available Wikipedia pages as references.

3.1.1. Distribution of research areas in Zooniverse projects

A first study targeted the (inter-)disciplinary nature of CS projects using a sample of 218 projects from the CS platform Zooniverse[3]. Zooniverse projects cover a wide variety of disciplines yet are relatively homogeneous in that they are based on a contributory and crowd-sourcing approach, in which volunteers are provided with data in online repositories. The task of the volunteers consists of data classification and analysis. The online representation of projects on the online platform is also relatively homogeneous.

Based on the available project descriptions in conjunction with the ESA approach with Wikipedia reference pages, research areas (RAs) were assigned to each project. We found that the most predominant RAs were *Biodiversity and Conservation* (80 projects), *Environmental Science and Ecology* (50 projects) and *Remote Sensing* (44 projects - see Figure 4 for more details). Since our approach allows for multiple RAs being assigned to each project, these are not necessarily distinct projects.

Our data analysis shows that 71 projects had only one RA assigned, 37 projects have two RAs and 37 projects have three RAs assigned to them. The two projects with the most RAs are *League of Nations in the Digital Age* and *Lakeside Dark Data* which have 18 and 22 RAs respectively.

[3] https://zooniverse.org

The RAs are divided into 5 categories – *Arts and Humanities*, *Life Sciences and Biomedicine*, *Physical Science*, *Social Science* and *Technology* – so we also have information on its distribution. *Life Sciences and Biomedicine* ranks highest with 165 projects, *Technology* comes second with 86 projects, *Physical Science comes* third with 53 projects, then *Social Sciences* with 38 projects and lastly, *Arts and Humanities* with 27 projects. Most projects were assigned RAs from one singular category (121 projects), but we also found projects with RAs from two (61 projects), three (23 projects), four (8 projects) and five (5 projects) categories.

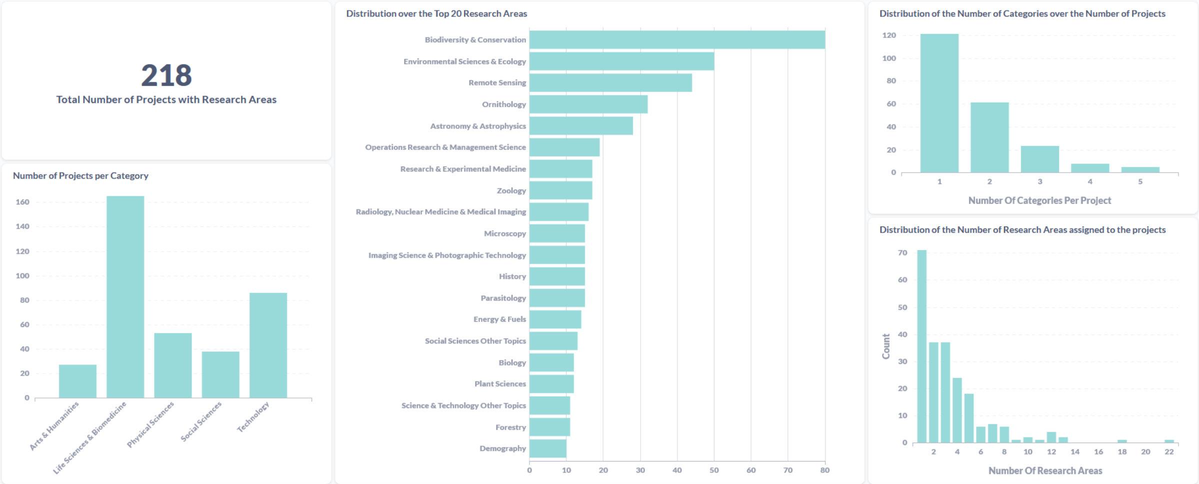


Figure 4: Dashboard on Zooniverse projects' research area assignments

Regarding RAs, we see a prevalence of multi-disciplinarity (average of 3.34 RAs per project) in a sample of 218 CS projects taken from the Zooniverse platform. The dominant RAs are *Biodiversity and Conservation* and *Environmental Sciences and Ecology* followed by *Remote Sensing*, *Ornithology* and *Astronomy and Astrophysics*.

3.1.2. Interdependence of research areas and SDGs

Several contributions to recent research on Citizen Science address the interconnection of CS projects with Sustainable Development Goals or SDGs (see Fraisl et al., 2020; Fritz et al., 2019; Moczek et al., 2021). This synergy between CS and the pursuit of SDGs is certainly positive and desirable, yet it can be problematic if used as a generalised criterion for judging the quality and relevance of CS activities since it "discriminates" certain well-established fields of CS in terms of research areas (RAs). A typical suspect for such a negative effect is the field of *Astronomy and Astrophysics* that was one of the foundational areas of Citizen Science. Based on our data and analysis approaches, we have calculated the "resonance" of RAs with SDGs as a basis for better informed judgement on these dependencies. These results have been presented at the ECSA conference 2022 (Hoppe et al., 2022).

For the analysis of associations between RAs and SDGs, we used a sample of 208 CS projects from different platforms. The ESA method was used in the same as described to automatically assign the association of projects (by their textual descriptions) to RAs and SDGs. Regarding the SDGs, we relied on the existing Wikipedia pages available for each SDG. However, to avoid a too strong dependency on the automatic method, we also used human coding to provide such an assignment (for SDGs). On this basis, we calculated the relative overlap between each combination of RA and SDG using the Jaccard measure of similarity (Figure 5).

$$Sim(X,Y) = \frac{(Number\ of\ projects\ associated\ with\ X\ \mathbf{and}\ Y)}{(Number\ of\ projects\ associated\ with\ X\ \mathbf{or}\ Y)}$$

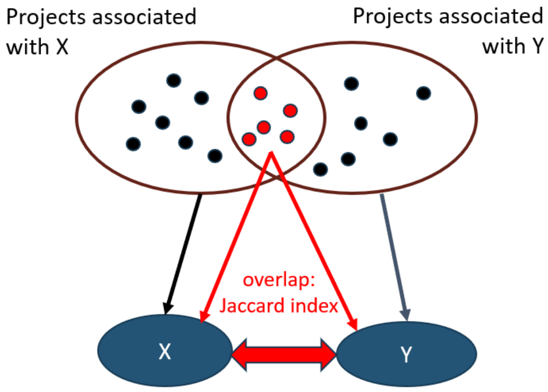


Figure 5: Jaccard measure of similarity based on assignments of descriptors X and Y to projects

We interpret this similarity measure as the degree of "resonance" between a given RA and SDG based on their co-occurrence for all projects in the sample. As argued in D1.2 (section 7.3), this is an empirical measure that is induced by the observed practice of CS projects. Table 4 provides an overview of aggregated similarities with all SDGs as well as specifically with SDG 4 ("Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all") for the top 10 RAs that were most frequent in our sample. SDG 4 has a particular relevance in this context since it may still "resonate" with areas of science that contribute to education without directly supporting "planetary health" and sustainability of natural resources. As to be expected, *Education and Educational Research* scores much higher on SDG 4 than on all SDGs on average. The RAs with the highest resonance levels are no surprise, as is the relative low score of *Astronomy and Astrophysics*. The same holds for *Arts and Humanities*. It is, however, surprising that *Ornithology* scored even lower than *Astronomy and Astrophysics*. A possible explanation that was corroborated by discussions during ECSA conference is that *Ornithology* projects are often very much focused on the continuous observation of certain species in in specific habitats without contextualising these observations in the light of sustainability and nature conservation in general.

Table 4: Dashboard on Zooniverse projects' research area assignments

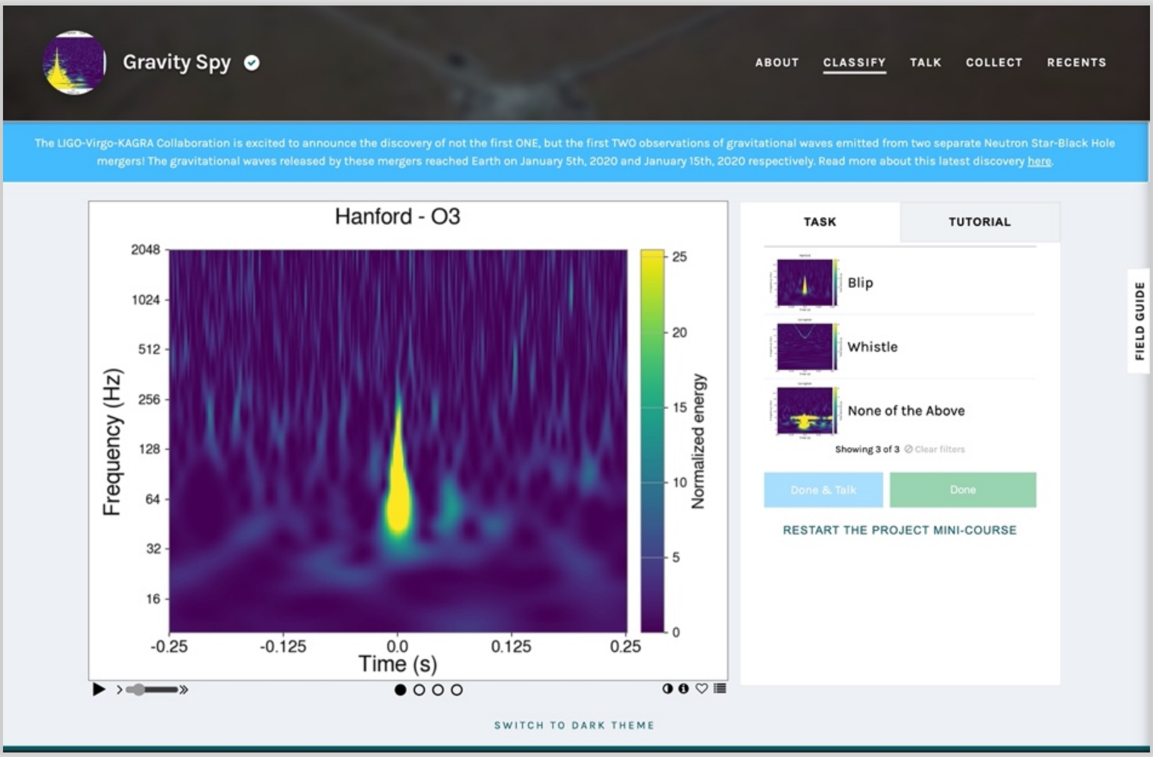
Research Area	Average resonance with all SDGs in %	Resonance with SDG 4 (Qual. Edu.) in %
Environmental Science & Ecology	16.1	17.8
Biodiversity & Conservation	13.5	16.2
Life Sciences & Biomedicine	11.2	20.7
Education & Ed. Research	7.9	26.5
Computer Science	4.9	5.4
Astronomy & Astrophysics	3.9	5.3
Ornithology	3.8	2.4
Arts & Humanities	3.7	9.2

3.2. Participation and motivation

As described above, one of the main research foci of this project is the analysis of volunteer' participation, engagement and motivation. Analyses in this respect can primarily be classified as *micro*-level analyses (see section 1), as they often depend on individual trace data on the project level. However, the developed techniques and steps of analysis can then be expanded to bigger samples involving more projects or platforms, thus moving from the *micro*- to the *meso*-level of analysis. In the following, we will describe such a case study, where we started with a sample of CS projects hosted on Zooniverse.

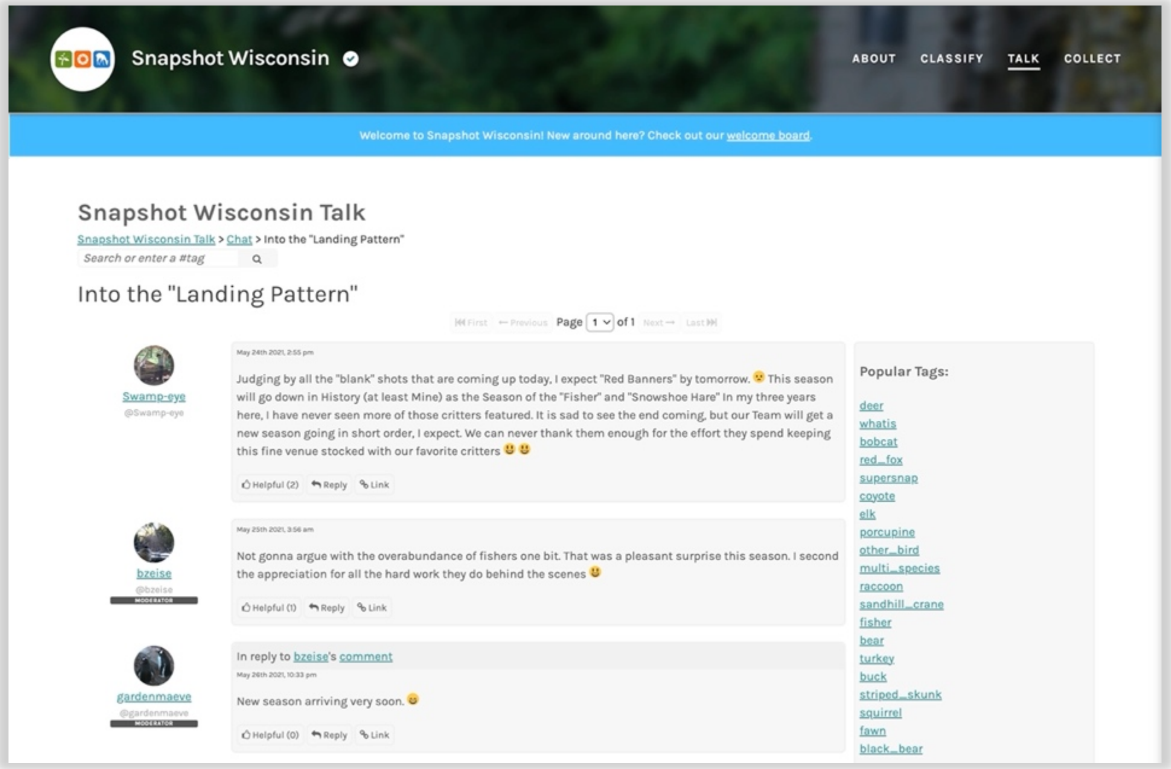
3.2.1. Contributory projects on Zooniverse

Online web portals like Zooniverse are of particular interest for our analyses, as they provide an open space for volunteers to engage in citizen science and collaborate with professional scientists (see Michalak, 2015). Zooniverse offers a vast array of different citizen science projects, ranging from astronomy to literature and ornithology. In all these crowdsourced projects, volunteers primarily contribute by classifying, annotating, or categorising data. Figure 6 shows such a typical classification task for the Gravity Spy project, where volunteers are asked to classify glitches in spectrograms from a gravitational wave detector. They can see the spectrogram that is to be classified, as well as different choices for the classification on the right. Other projects hosted on Zooniverse follow a similar structure, with data being "crowdsourced" by volunteers who contribute by working with data they are provided. As volunteers are typically not involved with other parts of the scientific process (i.e., collecting data in the field, formulating the research hypotheses), these projects can be labelled *contributory* projects (Bonney et al., 2009).



The LIGO-Virgo-KAGRA Collaboration is excited to announce the discovery of not the first ONE, but the first TWO observations of gravitational waves emitted from two separate Neutron Star-Black Hole mergers! The gravitational waves released by these mergers reached Earth on January 5th, 2020 and January 15th, 2020 respectively. Read more about this latest discovery here.

Figure 6: Example of a classification task for the Gravity Spy project



Welcome to Snapshot Wisconsin! New around here? Check out our welcome board.

Snapshot Wisconsin Talk

Snapshot Wisconsin Talk > Chat > Into the "Landing Pattern"

Into the "Landing Pattern"

Swamp-eye
@Swamp-eye
May 24th 2021, 2:55 pm
Judging by all the "blank" shots that are coming up today, I expect "Red Banners" by tomorrow. 😮 This season will go down in History (at least Mine) as the Season of the "Fisher" and "Snowshoe Hare" In my three years here, I have never seen more of those critters featured. It is sad to see the end coming, but our Team will get a new season going in short order, I expect. We can never thank them enough for the effort they spend keeping this fine venue stocked with our favorite critters 😃 😃

bzeise
@bzeise
MODERATOR
May 25th 2021, 3:56 am
Not gonna argue with the overabundance of fishers one bit. That was a pleasant surprise this season. I second the appreciation for all the hard work they do behind the scenes 😃

gardenmaeve
@gardenmaeve
MODERATOR
In reply to bzeise's comment
May 26th 2021, 10:33 pm
New season arriving very soon. 😄

Popular Tags:

deer
whatis
bobcat
red_fox
supersnap
coyote
elk
porcupine
other_bird
multi_species
raccoon
sandhill_crane
fisher
bear
turkey
buck
striped_skunk
squirrel
fawn
black_bear

Figure 7: Discussion forum of the Snapshot Wisconsin project on Zooniverse

In addition to the classification tasks, volunteers have the possibility to interact with each other or with members of the project team (i.e., scientists) through an integrated discussion forum ("Talk" page in the top menu). It consists of different forums (i.e., Chat, Science, etc.) in which users can create discussions and interact with each other. Figure 7 shows such a discussion. Each user is presented with their username, profile picture, and user role. The user roles can be assigned by the project owners or moderators, and the default role for users is *volunteer* (i.e., when no role is being explicitly shown). Other visible roles are *moderator*, *team*, *researcher,* or *translator*. The discussion forum provides a particularly relevant space for interaction to occur not only between volunteers, but also between volunteers and other user groups such as scientists or moderators. For example, volunteers might create discussions to seek content-related help by scientists or moderators might disseminate information and guidance regarding the task. Recent research on Zooniverse discussion forums corroborated this, by showing that moderators and highly active volunteers do indeed create new discussions, while scientists are often brought into discussions (e.g., by explicitly mentioning them using the forum's built-in function, see Rohden et al., 2019).

3.2.2. Goals and indicators

For our research on participation and motivation, crowdsourced CS project, like the ones on Zooniverse where volunteers merely take a *contributory* role (Bonney et al., 2009), are of particular interest. Recent research is primarily concerned with volunteer engagement and how it can be sustained to increase data volume and quality, as high attrition rates have already revealed that volunteers participate in these forums to varying degrees and that their behaviour is associated with the different roles adopted (Rohden et al., 2019). In this case study, our goal is to extend these findings, specifically by considering the working relationship between volunteers, scientists and other stakeholders within these projects. To this end, we are also interested in the motives and benefits that volunteers gain from participating, as these are not necessarily evident at first glance. We approach this by combining descriptive analyses and social network analyses of forum data, and by considering several quantitative indicators, extending earlier findings from a study of the Chimp & See project as a single case (see Amarasinghe et al., 2021).

The first and straightforward indicator when analysing the data is the actual discussion volume, i.e., the number of comments with respect to different user roles. Combining this with temporal aspects such as the time a comment was created can offer insights regarding participation patterns and involvement at certain points in the project's lifespan. By paying particular attention to the user roles, we can also detect any role changes and possibly changed participation behaviour during the project, as well as relate behaviour to user roles. Using SNA, we can extend this and consider the individual "importance" of users, i.e., by counting the number of connections each user has (degree), as well as ingoing directions (in-degree) and outgoing directions (out-degree). Other topological measures like the reciprocity (ratio of the number of edges pointing in both directions to the total number of edges in the graph) can help determine aspects like the mutuality of the discussions.

3.2.3. Data collection, sampling and processing

To collect the data, we extracted comments from Zooniverse using their API. Focusing on actual interactions, we only considered discussions with at least 2 comments. Thus, we extracted 2,049,646 comments from 703,139 discussions of 367 projects. An overview of our approach can be seen in Figure 8. The time of the collected comments ranges from the beginning of the respective projects until the time of our collection, which occurred in December 2021. On average, we have comments representing a timeframe of 6 years.

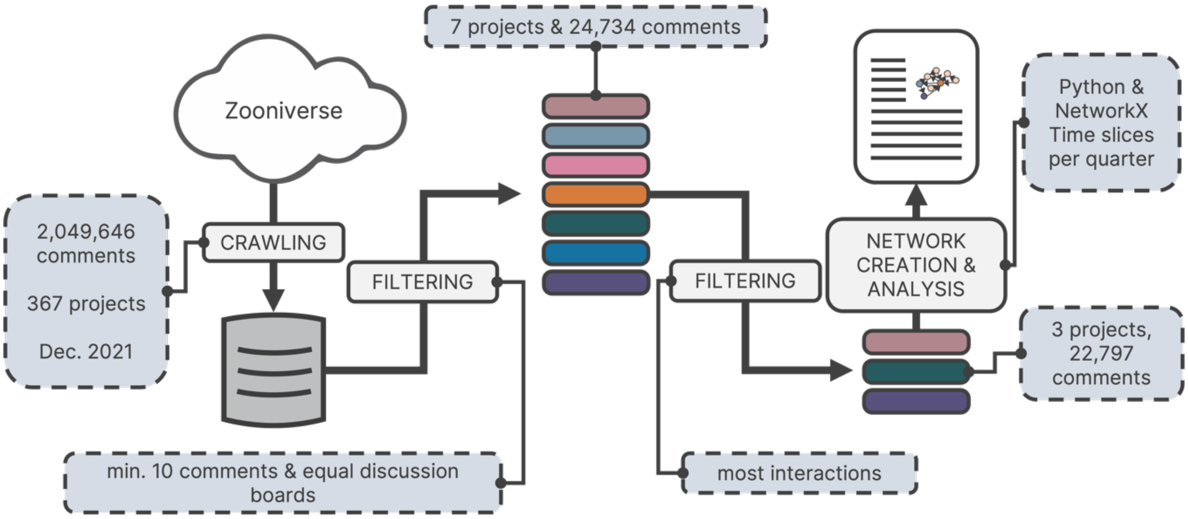
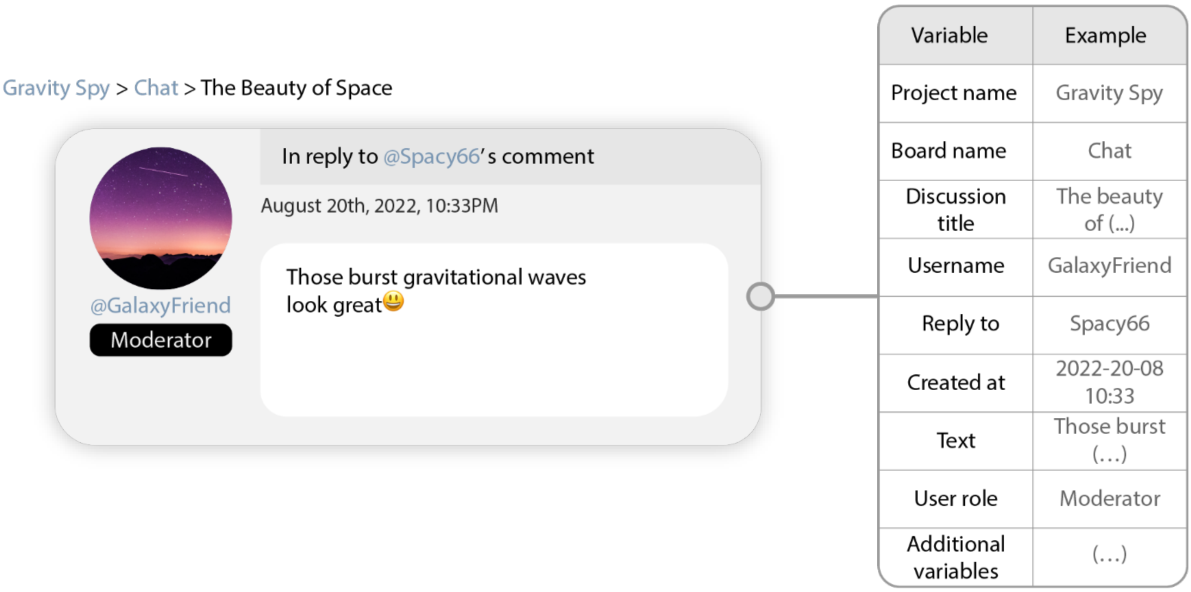

7 projects & 24,734 comments
Zooniverse
Python & NetworkX Time slices per quarter
2,049,646 comments
367 projects
Dec. 2021
CRAWLING
FILTERING
FILTERING
NETWORK CREATION & ANALYSIS
3 projects, 22,797 comments
min. 10 comments & equal discussion boards
most interactions

Figure 8: Overview of our analysis chain

The individual comments are the basic data representation in our sample. For each comment, we have multiple data fields, as can be seen in Figure 9. As was explained previously, several distinct user roles are visible within the forum data. When checking the data, we observed that many users have multiple roles at the same time, e.g., (*team* and *scientist*).



Gravity Spy > Chat > The Beauty of Space

@GalaxyFriend

Moderator

In reply to @Spacy66's comment

August 20th, 2022, 10:33PM

Those burst gravitational waves look great😃

Variable	Example
Project name	Gravity Spy
Board name	Chat
Discussion title	The beauty of (...)
Username	GalaxyFriend
Reply to	Spacy66
Created at	2022-20-08 10:33
Text	Those burst (…)
User role	Moderator
Additional variables	(…)

Figure 9: Example comment and respective data fields

In these cases, we defined absorption rules to reduce these multiple roles to a singular role. We defined the absorption rules in such a way that the most relevant role is kept. In certain cases, we allowed multiple roles to persist, as they might yield additional informational value. For example, if a user is a *scientist* and a *moderator*, we defined their new role as a *scientist-moderator*. Similarly, users who appeared as *volunteers* during their individual history become *volunteer-moderators* when their role changed during their time in the project.

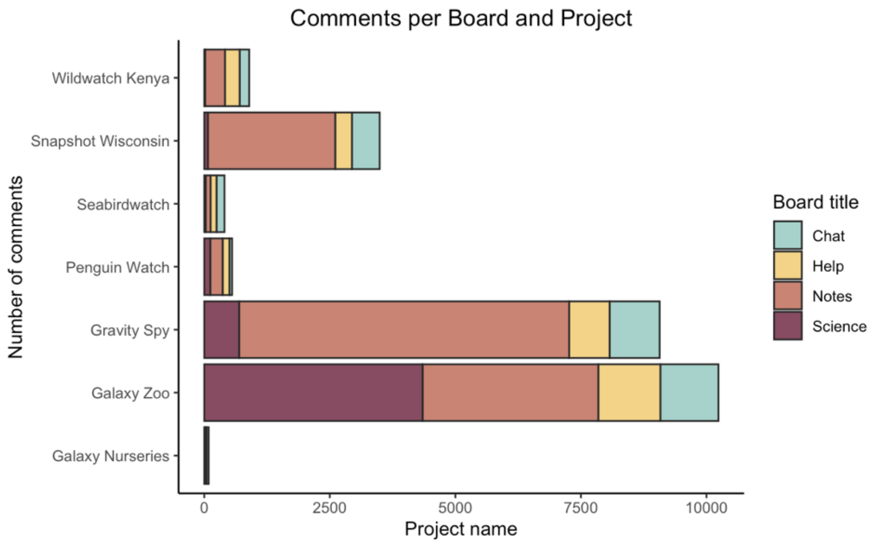


Figure 10: Distribution of comments per board and project

The dataset comprises discussions occurring in distinctive boards (i.e., sub-forums such as *Chat* or Science). The sub-forums serve as a categorisation mechanism by Zooniverse, as for example, it is encouraged to post scientific questions in the Science board. To allow for comparability, we limited our sample to projects which have the same boards, namely Chat, Science, Help and Notes. In these boards, we expect to see the most relevant discussions between the different users and user groups. Additionally, we limited our sample to discussions with at least 10 comments, as those with 2 comments most likely yield little informative value regarding interactions. According to these criteria, we limited our original sample of 367 projects to 7 projects, now comprising 24,734 comments. Figure 10 shows the distribution of our sample per board and project. We further limited our sample due to significant fluctuations in the number of comments to the three projects Galaxy Zoo, Gravity Spy and Snapshot Wisconsin.

3.2.4. Conceptualisation and extraction of networks

In section 3.2.2, we described how the transfer of data into network structures can offer additional insights that go beyond descriptive analyses. For this, we first conceptualised our network structure. Generally, online forum data contains an inherent relational structure that we can use to construct directed networks from. As demonstrated in related and earlier work (Amarasinghe et al., 2021; Hecking et al., 2017), we defined a user network with two relation types, the *reply* and *comment* relation: If user *u* comments on another user's (*v*'s) post, we create an edge (*u, v*) between them, and when *u* replies to a post by user *v*, we also create an edge (*u*, *v*). By using such a conceptualisation, we can depict user interactions by having the nodes represent the individual users, and the edges represent actual comments that connect them.

Although we have information on the time of creation for each comment, a simple network created from these comments does not necessarily reflect any temporal information, as it typically includes all comments (i.e., edges) irrespective of the time they were created. Therefore, we created multiple time slices, each reflecting a quarter of a year. Specifically, we first clustered the comments based on their time of creation, and then generated distinct networks for each time slice.

3.2.5. Results and interpretation

In the first step, we were interested in general patterns of participation. Related research on discussion forums in MOOCs (see communities of practice, Sarirete & Brahimi, 2014) already showed that most users do not actively participate in the discourse, and users that are highly active are in the minority. Thus, we calculated the top 5% of users, and examined their participation behaviour. Confirming related research, our analyses showed that those top 5% of users (*N* = 91) are indeed responsible for 75% of the total comments – meaning that the other 1,733 users merely created a quarter of the general discourse. In the next step, we were interested in how the participation in the forums changes over time.

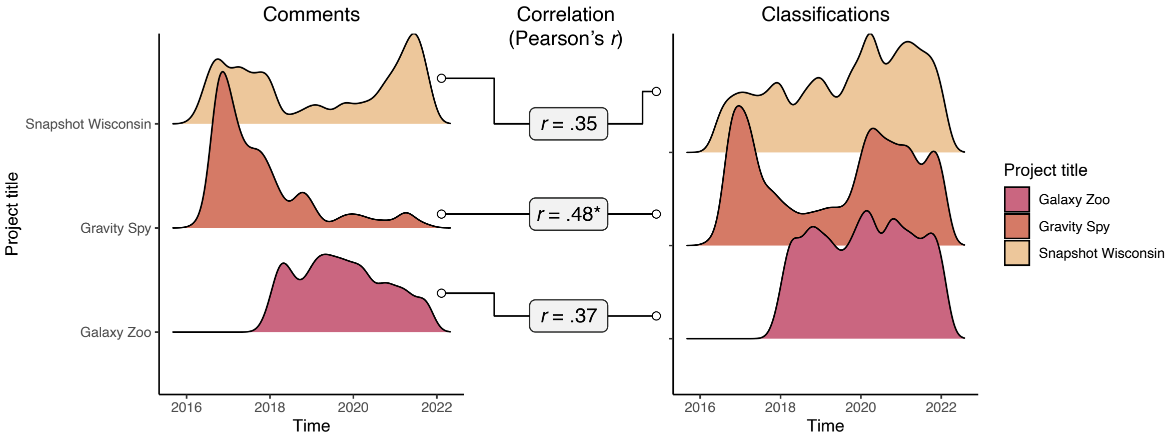


Figure 11: Comments and participation over time. Significant correlations are indicated by a star (*)

As it can be seen in Figure 11, there are considerable differences between the analysed projects. While participation activity and classifications generally appear to correlate, we see decreasing forum activity (i.e., comments) for *Gravity Spy* and *Galaxy Zoo*. However, for *Snapshot Wisconsin* this is not the case, as forum activity as well as classifications steadily increase. Interestingly, we see that for all projects, forum activity is highest during the beginning. This mirrors earlier findings (Amarasinghe et al., 2021) and can likely be attributed to increased help-seeking behaviour by the users and consequently, discussion activity in the forum. Thus, particularly in the beginning of the projects, the forum appears to be fostering interactions between the users. Additionally, a peak in productivity can be observed during the beginning of 2020 – likely being caused by the global pandemic, as many individuals had to stay at home and may have been driven to participate in such projects.

In section 3.2.2, we described how we can use topological network measures such as the degree or the reciprocity to examine interaction between users. To this end, we calculated the average degree (mean number of connections/comments per user) per time slice, as well as the corresponding reciprocity of the graph.

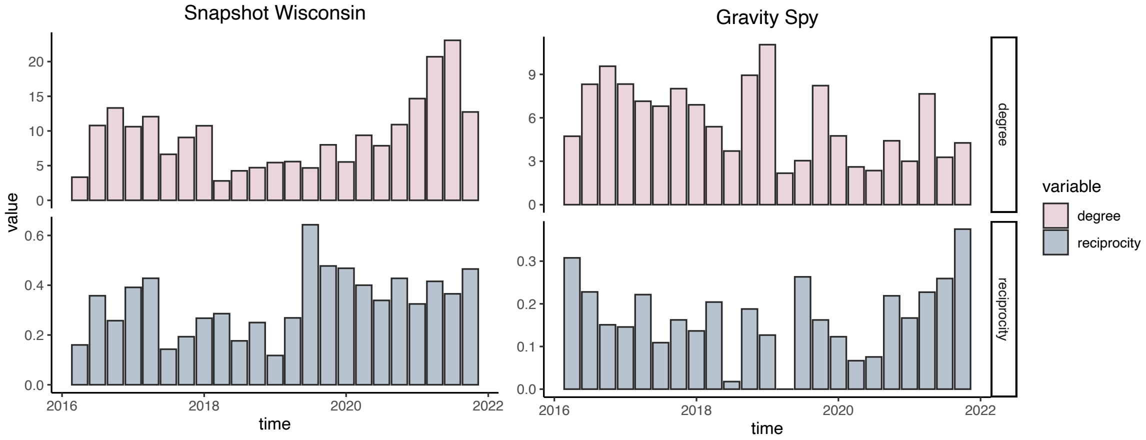
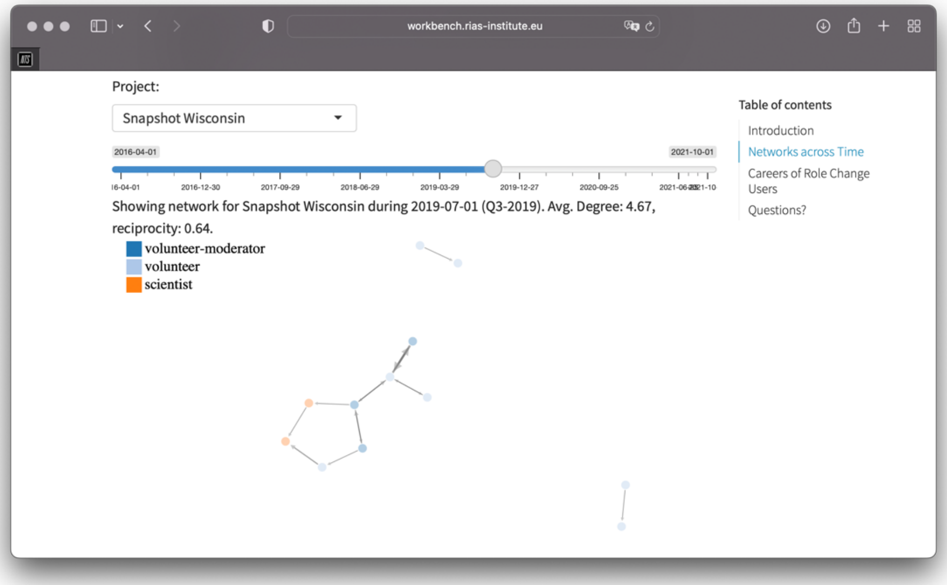


Figure 12: Avg. degree and reciprocity across time for Snapshot Wisconsin and Gravity Spy

As can be seen in Figure 12, on average, users become more central in the *Snapshot Wisconsin* project, while in *Gravity Spy*, they become less central. For both projects, reciprocity appears to increase, which indicates that the discussions tend to become more reciprocal. This can be interpreted as a sign for increased mutual engagement within the forum by the volunteers. Interestingly, during times of little interactions (and thus smaller degrees), there are cases where reciprocity is particularly high. An example to for this can be seen for the *Snapshot Wisconsin* project during mid-2019, where we see high reciprocity paired with a low avg. degree. Situations like this can be easily examined using the corresponding Quarto notebook we created, as it allows for an interactive generation and exploration of network visualisations based on user input. Figure 13 shows a screenshot of this notebook and how the corresponding network visualisation for the above-described situation looks like. As it can be seen, the thick edges with two arrows (i.e., bidirectional edges) indicate many mutual interactions between the connected users. The Quarto notebook (see also annex 6.2) was used in an interactive workshop during the ECSA conference, which will be described in more detail in section 4.



Project:

Snapshot Wisconsin ▾

Showing network for Snapshot Wisconsin during 2019-07-01 (Q3-2019). Avg. Degree: 4.67, reciprocity: 0.64.

- volunteer-moderator
- volunteer
- scientist

Table of contents

- Introduction
- Networks across Time
- Careers of Role Change Users
- Questions?

Figure 13: Network visualisation for Snapshot Wisconsin during Q3-2019 as shown by our interactive Quarto notebook

Of particular interest for our analyses were also the different user groups and how they interact with each other, as indicated by the different roles the users publicly display in the forum. Across the whole sample, with 76%, most comments come from *volunteers*, followed by *volunteer-moderators* with 13% and then *moderators* with 8%. The remaining 3% of comments are made by *scientists, scientist-moderators,* and *team*. We were further interested in how this share in the discussions behaves over time. Thus, we examined this across the whole time of our project, specifically based on our earlier described time slices. The results of this can be seen in Figure 14.

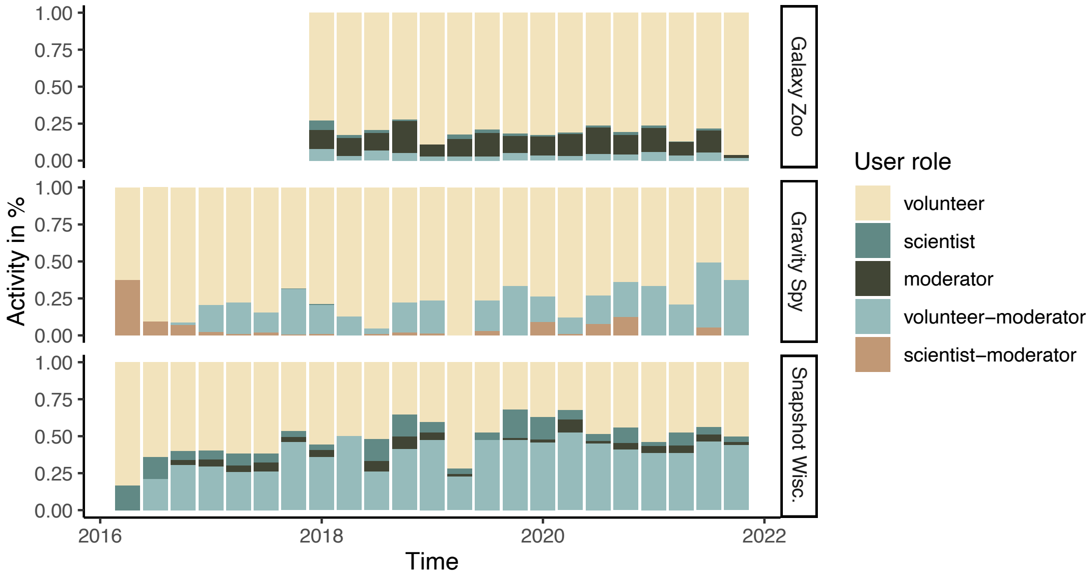


Figure 14: Share of comments per user role across time

With few exceptions, *volunteers* account for most comments throughout the project's duration. In the beginning, *scientists* and *scientist-moderators* maintain a high share in the discussions – which further underscores the point explained earlier, which is that during the beginning of projects, *scientists* contribute by giving help and possibly explaining the task. When considering the share of discussions by *volunteer-moderators*, a highly interesting picture emerges. *Volunteer-moderators*, who do not play a substantial role in the discussions during the beginning, steadily increase their share, until they nearly account for 50% of the discussions (see Figure 14). This, however, is not the case for *Galaxy Zoo*. To interpret this finding, we must consider what was described in section 3.2.3 regarding user roles and how they are assigned. On Zooniverse, users can change their roles, and become *volunteer-moderators*, for example. Specifically, this is the case for users who start to participate in the forum as *volunteers* and then, at some point, receive additional rights as *moderator*, thereby becoming *volunteer-moderators*. Such a role cannot be adopted by themselves, but is instead assigned from above (e.g., by *moderators*). As such, it can be seen as a reward, because it provides users with more rights and responsibilities. Thus, all users within our sample who are *volunteer-moderators* started as *volunteers* and changed their role at some point during the project, thereby getting promoted to a higher role. In our analysis, we found such 14 role changes.

A closer look revealed that 3 users changed their role within a single day, leading to their exclusion and a final sample of 11 users across three projects who changed their role. Although this number appears to be small (especially considering 1490 users in total), these 11 users account for 40% of the total comments and are therefore highly active. An example for this can be seen in Figure 15, where we see the relative share of users who changed their role in the discussions of the *Gravity Spy* project: The proportion of comments in the general discussions increases to 40% at times, and certain users "overtake" at some point during the project. For example, *user2*, who dominated most discourse during the early stages of the project got overtaken by *user6* around 2019, who then steadily increased their share. Thus, it appears that users who are highly engaged and even change their role to get promoted are part of this small user group which is responsible for most of the interactions within the Zooniverse forums. We also calculated group-comparisons to determine whether the differences in avg. degree between users who changed their role and those who participate a lot without changing their role are statistically significant, showing that this is indeed the case, and that the avg. degree is significantly higher for this user group (see Krukowski et al., 2022 for more details).

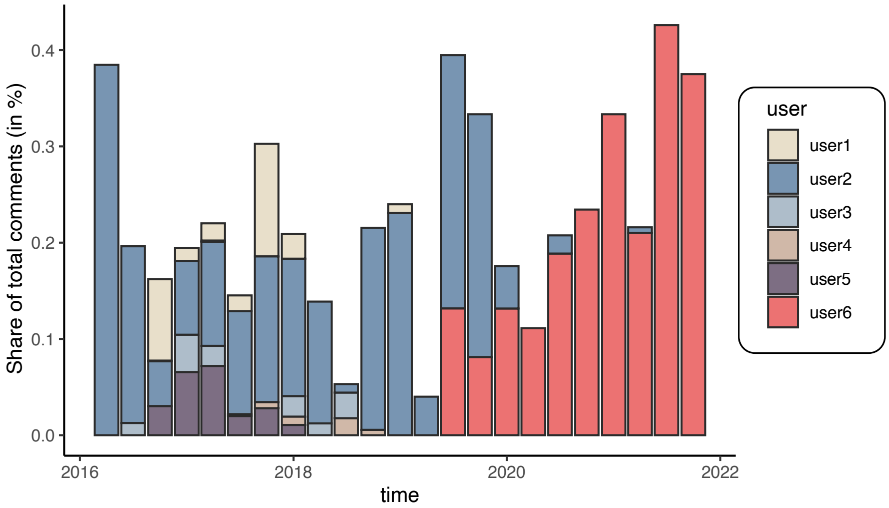


Figure 15: Share of role change users in the discussions for the Gravity Spy project

The fact that certain users changed their role and got promoted to a higher role not only indicates that such a promotion could potentially serve as an incentive mechanism, but it also poses the question as to how exactly such promotions occur, and also whether there are certain traces that can be seen in the behavioural data of these users. Thus, we considered the individual *trajectories* of these users and examined their specific development over time with reference to the time of the role change. For this, we primarily considered three indicators: The **degree rank**, the **in/out-degree ratio** and the **periodicity**. As could partly be seen earlier in Figure 16, the participation in the project forums is subject to fluctuations – in the example of *Snapshot Wisconsin*, participation volume increased towards the end, whereas the opposite is the case for *Gravity Spy*.

The participation of individual users is dependent on these fluctuations, because without much forum activity there is not much space for interactions of this user to occur. Therefore, we calculated ranks to account for this eliminating the effects of general high and low activity levels. The degree rank indicates the rank this user had in the respective time frame. Smaller values (i.e., 2) indicate higher ranks, and mean the user was in the top ranks concerning their degree in this time slice. For in/out-ratio, we consider absolute values, and shifts in the ratio of in- vs. out-degree can be highly insightful. A high in-degree means that the user received many comments (either by being replied to or by getting comments for a discussion they created), while a high out-degree means the opposite.

Particularly for behaviour like help-seeking or help-giving, a consideration of shifts in this respect can prove helpful. Lastly, we consider periodicity as a measure for continuous participation in the forum (c.f. Ponciano & Brasileiro, 2014) which indicates the mean days of absence (i.e., without forum interactions). In Figure 16, this can be seen for a user from the *Snapshot Wisconsin* project who changed their role and got promoted to a *volunteer-moderator*. For all measures, it can be seen, that the behaviour changed after the role change. The user consistently stayed in the top ranks regarding the degree, and steadily increased their degree as the general participation in the forum increased. While they received more comments *before* the role change (i.e., higher in-degree), this turned after the role change and the user gave more comments to other users. For periodicity, a less clear picture emerged, yet around the change, the periodicity was low, meaning the user was consistently participating in the forum.

The described case study shows how volunteers in online citizen science platforms, specifically Zooniverse, participate, engage and interact with each other. We were able to show that a small number of highly engaged users shaped most of the discourse in these forums, and that some of these users even changed their role and got promoted to a higher role. As such roles (i.e., *volunteer-moderator*) come with more rights and responsibility, promotions can be seen as an incentive for volunteers to participate and engage in scientific discourse. A closer look at their trajectories also shows that such a change is visible in their behavioural pattern – meaning that such information could potentially be used to incentivise participation. The study gives an idea about the possibilities of the different analysis techniques we developed, and it delineates how these techniques got more sophisticated over time, now allowing for a fine-grained analysis that could potentially be applied to any given CS project on Zooniverse.

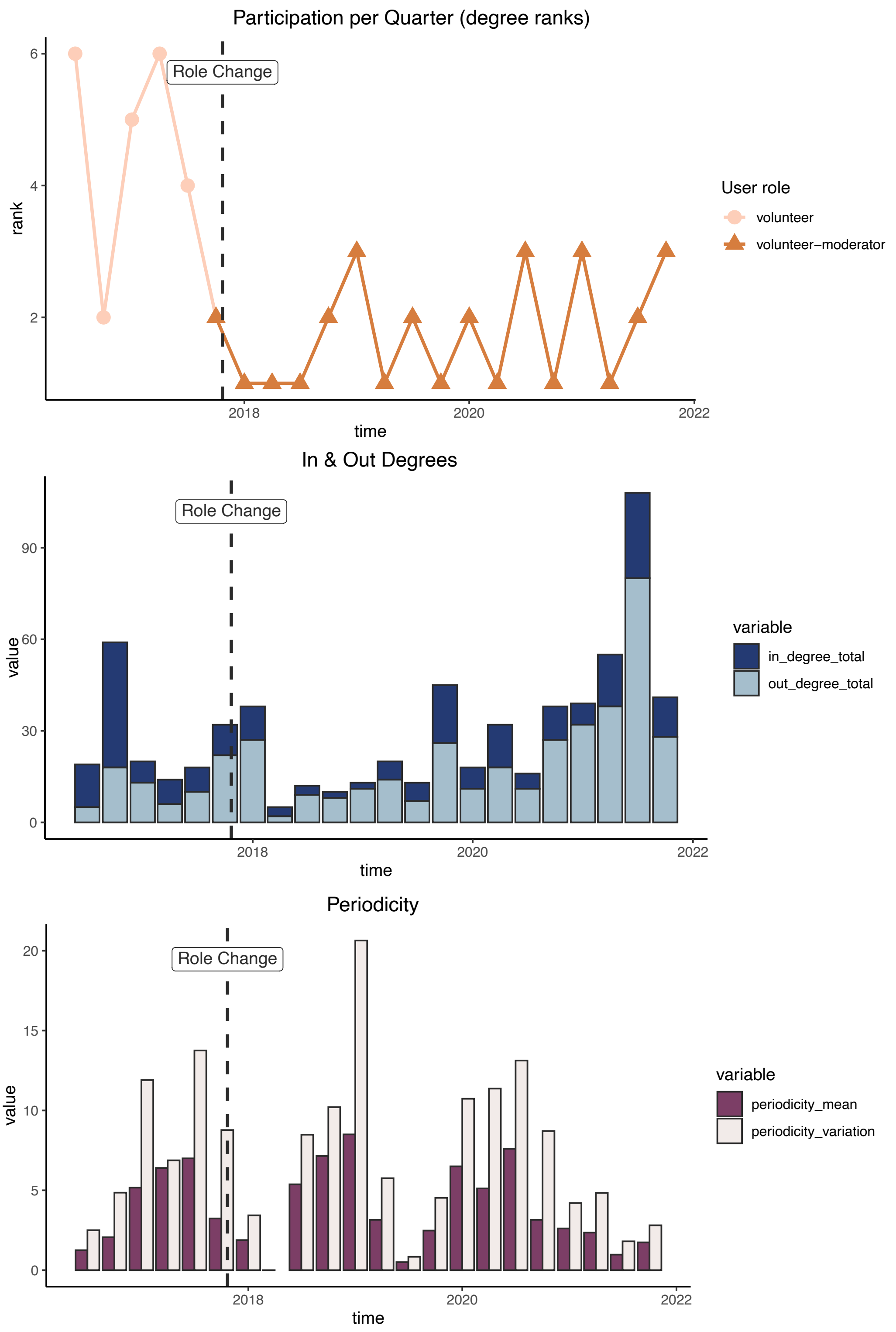


Figure 16: Individual trajectory with three indicators of a user who got promoted to volunteer-moderator

3.3. Analyses of Twitter data

In the context of CS Track, the analysis of Twitter data is an essential feature related to the macro level analysis (see section 1). It extends the scope of the project information contained in the database due to the presence of projects, organisations and individuals sharing information about Citizen Science on this social networking platform. The structure of Twitter allows for different types of analysis. These include content analyses based on hashtags, mentions or common words, which can be combined with deep learning techniques and sentiment analysis. All these methods, firstly proposed in D3.1 and described in D3.2, have a strong focus in text processing and text analysis. In addition, the evolution of methodology has led to a more diverse, complete, and complex set of processes including machine learning, API usage and exhaustive network analysis. In D3.2, we presented the deployment of the DASH dashboard, in which all the different analyses were explained alongside the visualisation of the results of those analyses.

These methods have proved themselves as valuable and of interest. In different case studies, we obtained different key findings in different aspects of the Citizen Science community. This findings in the studies about learning, SDGs, eHealth and climate change have turned into several publications in journals (De-Groot et al., 2022; Roldán-Álvarez, Martínez-Martínez, & Martín, 2021; Roldán-Álvarez, Martínez-Martínez, Martín, et al., 2021) and in the eMagazine of the project (Martínez Martínez et al., 2022a, 2022b, 2022c; Roldán-Álvarez & Martínez Martínez, 2021).

In the next sections, we go through the different methods accompanied by the key findings yielded.

3.3.1. Data analysis, text processing and text analysis

As stated in D3.2, the set of tweets that were used in the different analyses were extracted by means of the Lynguo tool[4]. This tool was designed to automatically collect tweets related to CS based on a set of keywords associated with Citizen Science, such as "Citizen Science" or "Citsci", and the translation of these words to other languages from the European Union. The harvesting of tweets started the 30th of September 2020 and is continuous, i.e., it is still ongoing. Currently, the total amount of collected tweets exceeds more than 700.000 tweets.

In these tweets, CS topics are discussed in general and in a variety of contexts. To isolate or extract different topics, we designed a filter based on keywords. Alongside the filtering based on keywords, we designed a function to remove tweets that could be identified as coming from bots based on "signal words" and behaviour over time. With this function, we can clear the noise inside the set of tweets to analyse. The remaining tweets were cleaned from stopwords and non-meaningful symbols.

In the sets of tweets that were selected through specific keyword filtering, we find discussions around the topic of our interest as a starting point for further analyses. For each set, we count the number of tweets, the number of users and the number of retweets given inside that topic. In all our studies we found that retweets are the biggest part of the data, a finding aligned with what has been already stated (Martínez Martínez et al., 2022b). The text analytics begin with the analysis of the hashtags, which is important since adding hashtags to the tweets is the easiest way to assign them to a specific topic. One of the main findings is that the most used hashtags are those related to Citizen Science. The most used one is always #CitizenScience in every analysed topic. When studying the SDGs (Sustainable Development Goals)[5], we found that climate change is the second most used hashtag. In Figure 17, we can see that most of the hashtags are related to climate change.

[4] https://lynguo.iic.uam.es/

[5] https://sdgs.un.org/goals

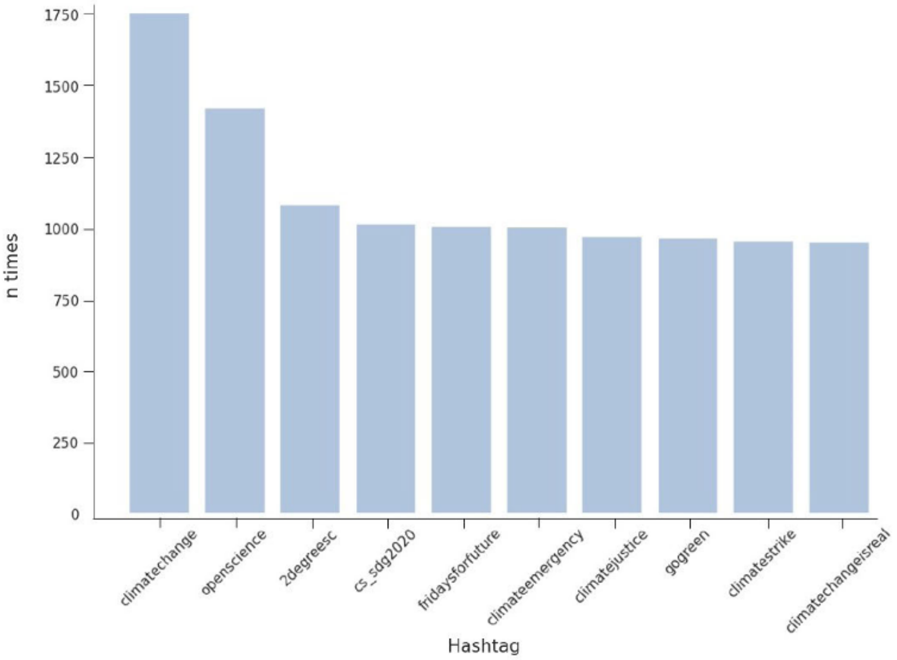


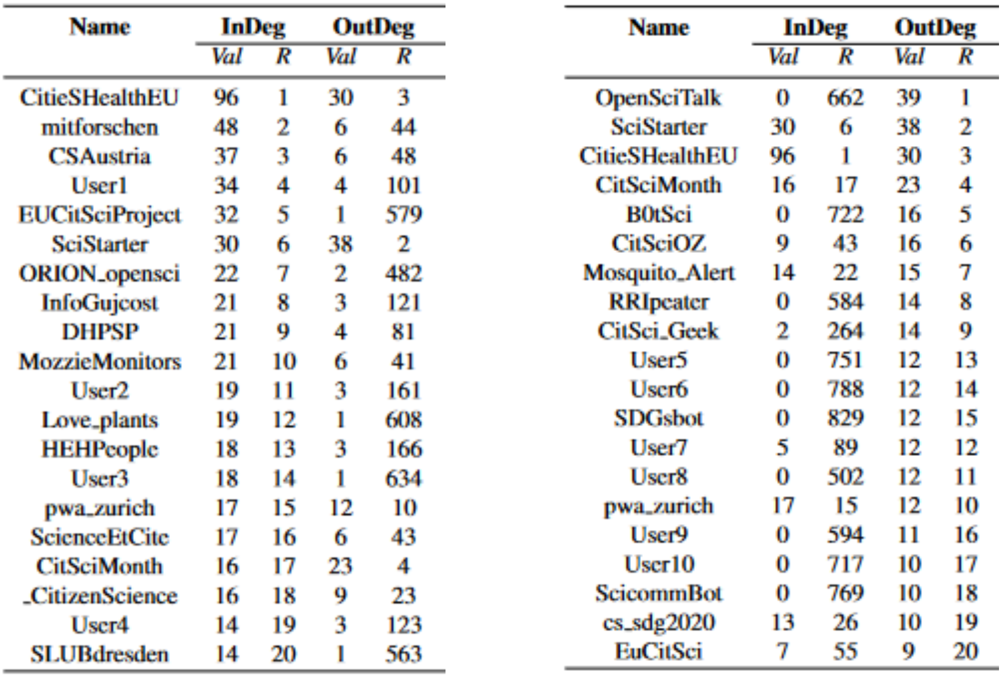
Figure 17: Most used hashtags in SDGs discussion

One interesting finding was, when studying the discussion about eHealth and learning, SDGs is the most used hashtag. This finding led us to believe that the SDGs are of high importance inside the Citizen Science community.

We followed a similar approach to analyse how the users mention each other. Instead of finding the hashtag symbol (#), we checked for the at symbol (@) followed by a username. This was only tested in relation to the eHealth topic, and we found that the platforms and projects are the most mentioned, for example, the CitSciOZ association (Australian Citizen Science Association). This finding is also aligned with what was stated by Mazumdar & Thakker (2020), which is that in the Citizen Science community in Twitter, the main part of the information disseminated are retweets and the remaining information tends to include replies to the users that wrote the retweeted tweet, which is seen through a mention of the original account. Projects and platforms are those that are most retweeted, and the users that mention them the most are the ones who support the original statement.

In the same line, one of the main analyses we always perform, is the analysis of the retweets. Retweets are one of the biggest parts of the data, as the connection between users (one retweets/shares what other user has posted) establishes a link between the members of the community which relate to a networked structure. This analysis of who are the most retweeted users and who retweets more is directly linked to the network construction, which will be properly explained in the next sections. As a preliminary explanation, when we connect two or more users via the retweet link, we are pointing this action in one direction. This direction establishes who retweets as well as who receives the retweet. In network analysis, we call the former the InDegree (InDeg, number of links pointing towards the user) and the latter OutDegree (OutDeg, number of links pointing outwards from the user) (cf. section 3.2.2). To provide more context on how we detect the retweets, we know that each retweet is labelled as “RT @username:”, making them very distinguishable from the rest of tweets. Besides counting the number of these elements, we also follow a similar method as the used one for hashtags and mentions. In this case, we search for the “RT @username:” pattern and are therefore able to analyse who receives the highest number of retweets. One of the key findings is that normally, the platforms and projects receive more retweets than the individuals. Also, when checking who retweets the most, the individuals appear to be particularly important. In Table 5, we see an example of this result. InDeg shows the number of retweets received, while OutDeg shows the number of retweets given. R tells the ranking in the set of users that retweet or receive retweets.

Table 5: Most retweeted accounts (left), most retweeting accounts (right)



Name	InDeg		OutDeg	
	Val	*R*	*Val*	*R*
CitieSHealthEU	96	1	30	3
mitforschen	48	2	6	44
CSAustria	37	3	6	48
User1	34	4	4	101
EUCitSciProject	32	5	1	579
SciStarter	30	6	38	2
ORION_opensci	22	7	2	482
InfoGujcost	21	8	3	121
DHPSP	21	9	4	81
MozzieMonitors	21	10	6	41
User2	19	11	3	161
Love_plants	19	12	1	608
HEHPeople	18	13	3	166
User3	18	14	1	634
pwa_zurich	17	15	12	10
ScienceEtCite	17	16	6	43
CitSciMonth	16	17	23	4
_CitizenScience	16	18	9	23
User4	14	19	3	123
SLUBdresden	14	20	1	563

Name	InDeg		OutDeg	
	Val	*R*	*Val*	*R*
OpenSciTalk	0	662	39	1
SciStarter	30	6	38	2
CitieSHealthEU	96	1	30	3
CitSciMonth	16	17	23	4
B0tSci	0	722	16	5
CitSciOZ	9	43	16	6
Mosquito_Alert	14	22	15	7
RRIpeater	0	584	14	8
CitSci_Geek	2	264	14	9
User5	0	751	12	13
User6	0	788	12	14
SDGsbot	0	829	12	15
User7	5	89	12	12
User8	0	502	12	11
pwa_zurich	17	15	12	10
User9	0	594	11	16
User10	0	717	10	17
ScicommBot	0	769	10	18
cs_sdg2020	13	26	10	19
EuCitSci	7	55	9	20

This led us to believe that, in general, platforms and projects create the content and receive a big number of retweets from the rest of users. It is known that the hashtags are important inside a tweet, but there are many more words inside the text. This is the reason why we checked which were the most used words inside the tweets. To perform this analysis, we removed the most common stopwords in several languages (English, Spanish, French, German, Dutch, Italian and Portuguese), we also removed symbols and punctuation to leave just words in the texts. We also removed every word preceded by # or @ to remove hashtags and mentions. Finally, we selected only those words that were nouns, verbs, and adjectives. The result is shown in Figure 18, where we see that Citizen, Science and climate are the most used ones. We also see some other words as support, join, research project and data.

The latest analyses in relation to the text analytics count most used words and extracting TF-IDF. With this technique, we tried to extract relevant words alongside those most used. The most common words inside the discussion around SDGs show that, as expected, Citizen and Science are the most used ones (Figure 18). Besides those two, we found words like climate, project, research, data, help or change to be relevant. However, we see a very different situation in Figure 19, as the most relevant words are not related to the most used ones.

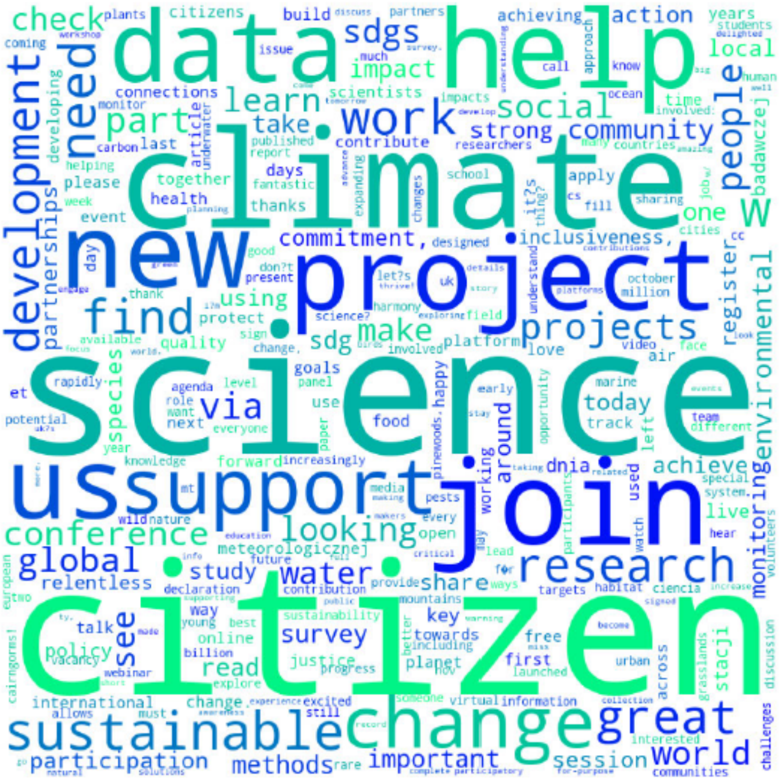


Figure 18: Most used words in SDGs conversations

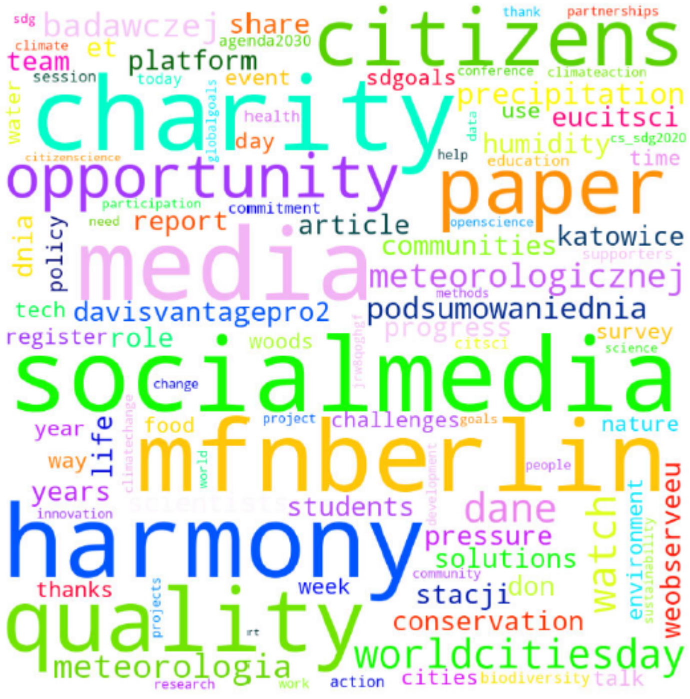


Figure 19: TF-IDF

3.3.2. Machine learning, BERT and LDA

In the line with text analytics, machine learning (ML) techniques can provide better and more detailed insight into the state of discussions. ML and deep learning approaches facilitate the generation of models for classification, prediction and pattern recognition with a high degree of accuracy, primarily depending on big volumes of data. In the analysis of social media interactions, ML techniques help to identify patterns and trends in datasets containing numerous objects, allowing us to perform different applications such as fake news recognition, sentiment analysis or topic discovering. In all of our studies, we have applied different ML approaches in order to discover the topics and sentiment in Twitter discussions in addition to more traditional algorithms such as LDA (Latent Dirichlet Allocation) or topic modelling (Blei et al., 2003) . Among the modern ML approaches, BERT (Bidirectional Encoder Representations from Transformers) has performed particularly well both in terms of results as well as the possibility to customise the parameters of the model and to finetune it with own texts.

LDA was used to analyse the topics of discussion inside the general SDGs related tweets. This algorithm checks the different texts and identifies patterns and relationships among text documents, clustering them in these topics according to their similarities. To obtain better results, we cleaned the documents by removing signs, symbols, punctuation and stopwords, this improves the performance of the model because the symbols used in Twitter as the URLs, hashtags and so on add noise to the texts and by removing them, we obtain plain texts more easily interpretable. With LDA, we can choose the maximum number of topics into which the tweets can be divided. By studying the topic coherence, we ended up with 17 topics of discussion.

For each topic, we gave them a name as a summary of what is discussed in them:

- Topic 1: WeObserveEu project
- Topic 2: Impact of technology in sustainable development
- Topic 3: The Citizen Science SDG conference
- Topic 4: Healthy planet in the future
- Topic 5: Online data more accessible
- Topic 6: Participation in sustainable development (water usage and climate change)
- Topic 7: Open Science
- Topic 8: Resolve climate change
- Topic 9: General SDGs discussion
- Topic 10: Use of technology to reach the SDGs
- Topic 11: Policies to be implemented around data, discussion in the Citizen Science SDG conference
- Topic 12: Youth actions to tackle climate change
- Topic 13: Impact of waste and use of geospatial data
- Topic 14: Bot tweeting about weather conditions in Katowice
- Topic 15: Reduce individual use of energy to reduce impact
- Topic 16: Bot tweeting about SDGs
- Topic 17: ObservaTree project and the warning about health of trees

Six out of the 17 topics were about climate change, with climate change being the most addressed topic in general as we can see in the results from all the different analyses. The complete analysis can be read in (Roldán-Álvarez, Martínez-Martínez, Martín, et al., 2021). In Figure 20, the different topics can be seen in an intertopic distance map, which shows the similarity between topics. In Figure 21, we see the most salient terms in the different topics.

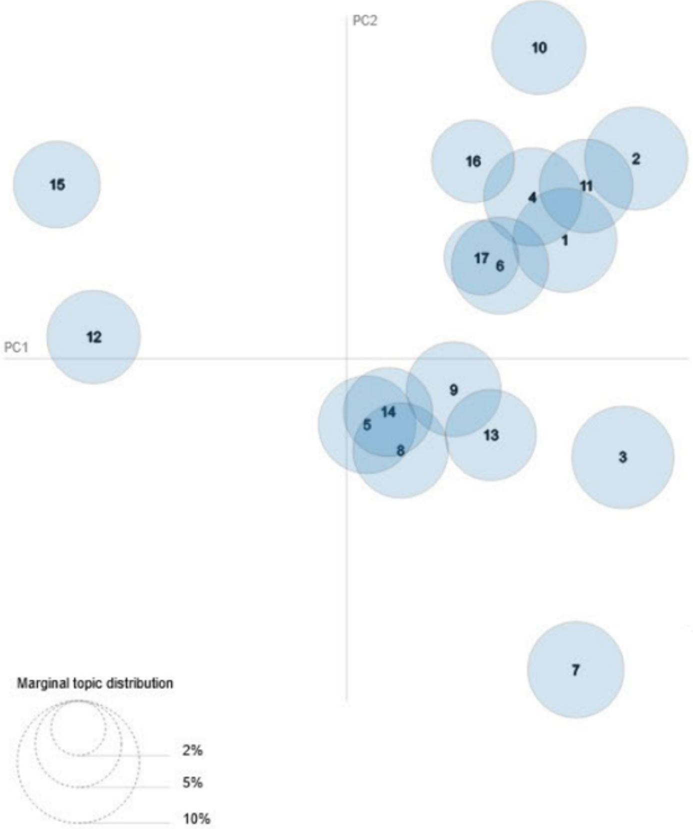
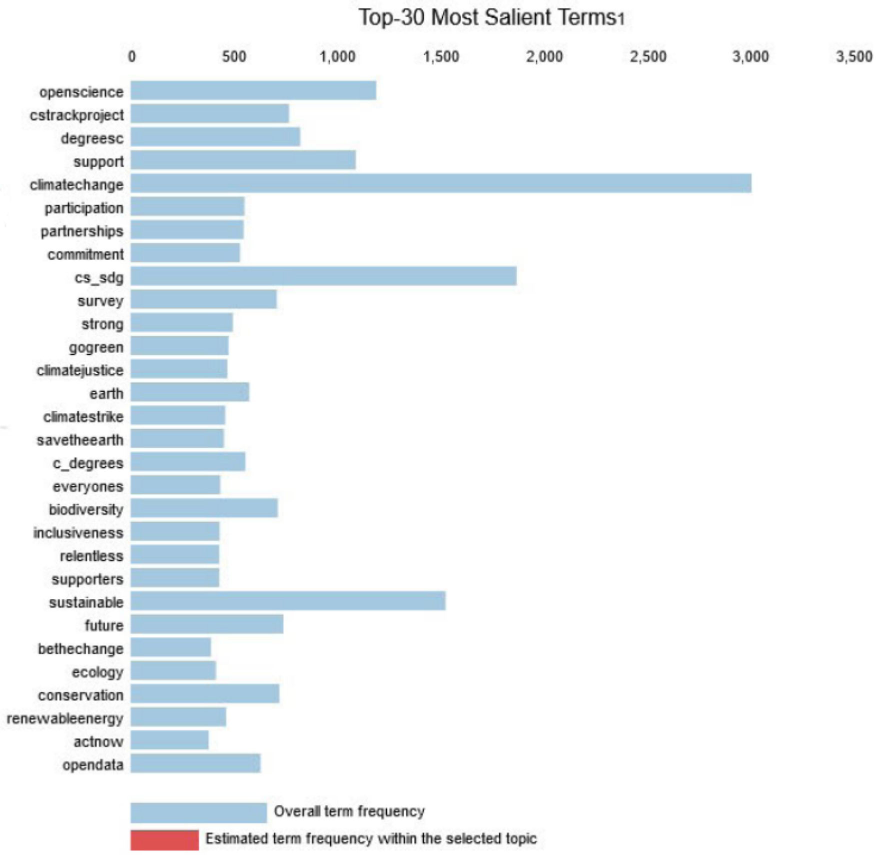


Figure 20: Intertopic distance map

In the same study, we applied a BERT classifier algorithm to the documents to classify them according to the SDG referred in the text. This classifier was trained to catalogue the tweets according to the SDG they address in the text. To train the classifier, we used 57,843 tweets about SDGs extracted from the Twitter API from May 1st, 2021 to May 15th, 2021 and used 80% of them for training and 20% for testing based on a random sampling. The overall score of classification was 0.82, which falls in the range between 0 and 1 (with 1 indicating perfect classification). As such, values of 0.82 represent a relatively high score with this technique (which has even been improved in recent works by extending the training process with new tweets).

Once we applied it to our tweets about SDGs, we found that the most addressed SDG is SDG 13: Climate action. This finding strengthens our previous belief that climate change is of great interest inside the Citizen Science community in Twitter, followed by SDG 11: Sustainable cities and communities, and SDG 10: Reduced inequalities. In Figure 22, we can see the distribution of tweets assigned to each SDG.



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 21: Most salient terms in SDGs discussion

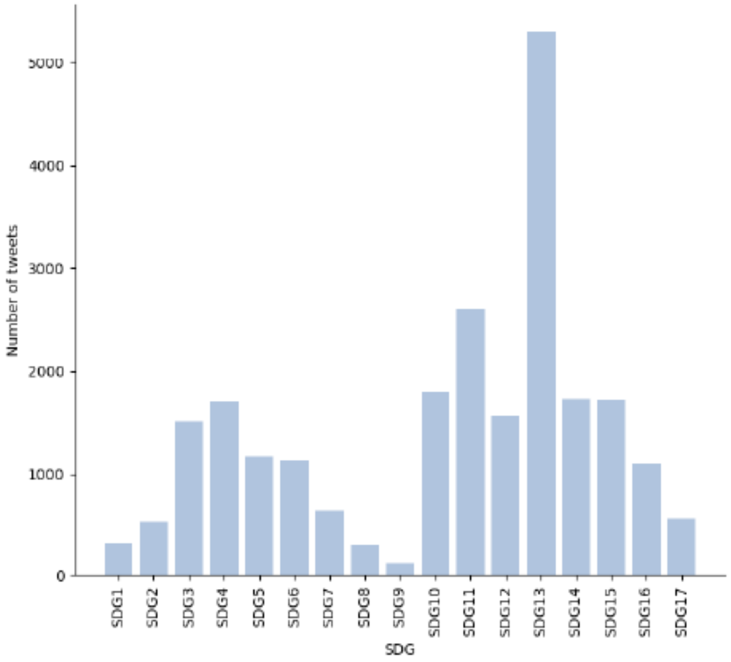


Figure 22: Tweets assigned to each SDG

In the following studies, BERT was used as the main algorithm for topic analysis and classification due to its versatility and good results. When studying the discussion around eHealth we applied a new version of BERTopic refined for social media analysis with our own tunings based on the HDBSCAN model (Campello et al., 2013). This algorithm is designed to improve the aggregation of similar topics into one only topic in a stable way. For this purpose, the algorithm checks the complete dataset and finds the similarities between elements, then aggregates those sharing similar values based in a statistical value named epsilon. The tuning also uses a UMAP model (McInnes et al., 2020) to strengthen the analysis of similarities between texts to aggregate them in topics. The texts are displayed in a graph structure and the algorithm tries to reduce the dimensionality (approaching the neighbours based in their likeness). Also, we used a sentence transformer model (Reimers & Gurevych, 2019) named "multi-qa-mpnet-base-dot-v1", a highly trained model with good results used to compute similarities in the meaning of sentences, helping the aggregation too.

In addition, we designed a script for downloading the tweets alongside the topic they were assigned to, a feature that it is not so well implemented in BERT, allowing us to perform further analyses. The results showed 19 topics, with "Mosquito Alert" being the most addressed, in which discussions evolved around health problems related to mosquitoes. The next important topics were about mental disorders, research on rare diseases and water sanitation. It was particularly surprising to find such a small number of tweets discussing COVID-19 inside Citizen Science. BERT also allows for a temporal analysis of the topics, letting us check how the popularity of these topics change over time. In Figure 23, we can see how they have evolved.

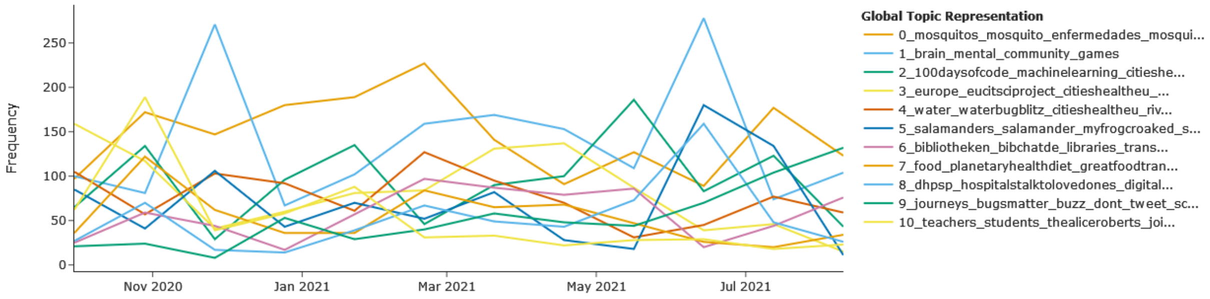


Figure 23: Evolution in time of the topics

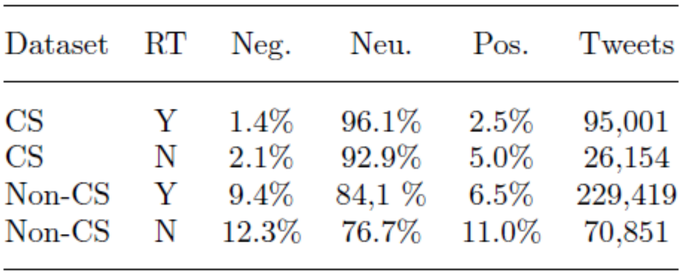
This analysis is especially relevant as it allows researchers and the audience to discover moments in time when these topics are highly discussed, due to the happening of events, conferences, etc. For example, we see a peak in Mosquito Alert in early March 2021 when the conference *Mosquito Control* was happening. Also, those high peaks of mental disorders could explain the happening of several webinars, mental health conferences in those dates such as: Mental Health in light of COVID-19 Virtual Summit (December 2020) and Mental Health and Human Resilience (May 2021).

One of our latest approaches was to use the BERT classifier to apply sentiment analysis. The sentiment analysis is a technique to classify text as positive, neutral, or negative, according to the words used in them, since the users who wrote them could be expressing emotions in the tweets. This analysis presents a new and interesting approach since we wanted to compare the sentiment in the tweets about climate change written by users in the CS community and tweets from outside the community.

To do this, we trained BERT with the sentiment140[6] dataset, which contains 1.6 million tweets classified according to the sentiment they express and that has been addressed in similar studies.

In Table 6, we can see the results of the sentiment analysis of the tweets about climate change. The results show that tweets on climate change inside and outside CS are eminently neutral, which was expected within the CS community, but this was not an expected result when analysing the sentiments outside CS. The table also contains results when counting the retweets, shown in the column RT when the value is Y (Yes). This means that every time a tweet is retweeted, this appears several times in our dataset and therefore it is given a value for sentiment. We discovered that there is a tendency for the most retweeted tweets to be neutral inside the CS Community, but outside the CS community, the most retweeted tweets are those showing sentiment, so the retweets add polarisation to the results.

Table 6: Sentiment predictions



Dataset	RT	Neg.	Neu.	Pos.	Tweets
CS	Y	1.4%	96.1%	2.5%	95,001
CS	N	2.1%	92.9%	5.0%	26,154
Non-CS	Y	9.4%	84,1 %	6.5%	229,419
Non-CS	N	12.3%	76.7%	11.0%	70,851

In our latest studies, besides downloading the information of the users, we downloaded the names of the accounts that certain users followed and were followed by. All this information is extracted and analysed always since we have access to the Twitter API and we abide by its rules, respecting the users ´privateness by not making their information publicly available but presenting the results to contribute to the public conversation. This data was used in the creation of ego-networks, a portion of the complete network focused on one local node and the nodes to whom this node is connected. This will be further explained in the next subsection about network analysis.

3.3.3. Network analysis

The relations between the users in Twitter (following), the retweets, the mentions, even the elements inside the tweets can be analysed in the form of networks. These interactions between the accounts or the elements establish a link between them, and we can approach the analysis of these links using social network analysis (SNA) with its various analytic measures (see D3.1). In the context of the analysis of social media, and more specifically in Twitter, the relations or ties of the underlying network can be of various types, including the follower relation, retweet connections or semantic links. In the first steps of the analysis, we used the data from Lynguo to analyse the networks of retweets, hashtags and mentions leading to the usage of data from the API to create more specific networks and comparisons between users creating ego-networks and the comparison between users outside and inside the CS community.

Initially, we created the network of retweets by connecting the user who retweeted the user who posted the tweet (A → B). Retweets, as it has been already stated, are of great importance since they consist of almost three-fourths of the data from the CS community. The retweeting relation allows us to understand who is potentially influential in the network since the content they are creating is

[6] http://help.sentiment140.com/home

shared and of interest to the rest of nodes. In our study about open learning, published in the ICALT 2021 conference (see Roldán-Álvarez, Martínez-Martínez, & Martín, 2021), we presented our first analysis of retweets. In this network, we can visualise the results of the analysis of retweets explained in 4.1.

In the next publications, we improved the analysis by adding more features to the network such as the weight in the edges according to the number of retweets given, the reduction of nodes in the visualisation using *k*-core algorithms and the calculation of communities. In the publication in IEEE Access about SDGs (see Roldán-Álvarez, Martínez-Martínez, Martín, et al., 2021), we implemented the use of the weighted edges and the trimming of the network with the goal to improve the statistical analysis in the network because now not only the nodes show the degree (number of RTs, mentions, use, etc.), but we also have edges with hierarchical values. The trimming of the network allows for a cleaner and more interpretable analysis and glance at the network, since we are not displaying thousands of nodes but those that are more connected. The visualisation can be seen in Figure 24.

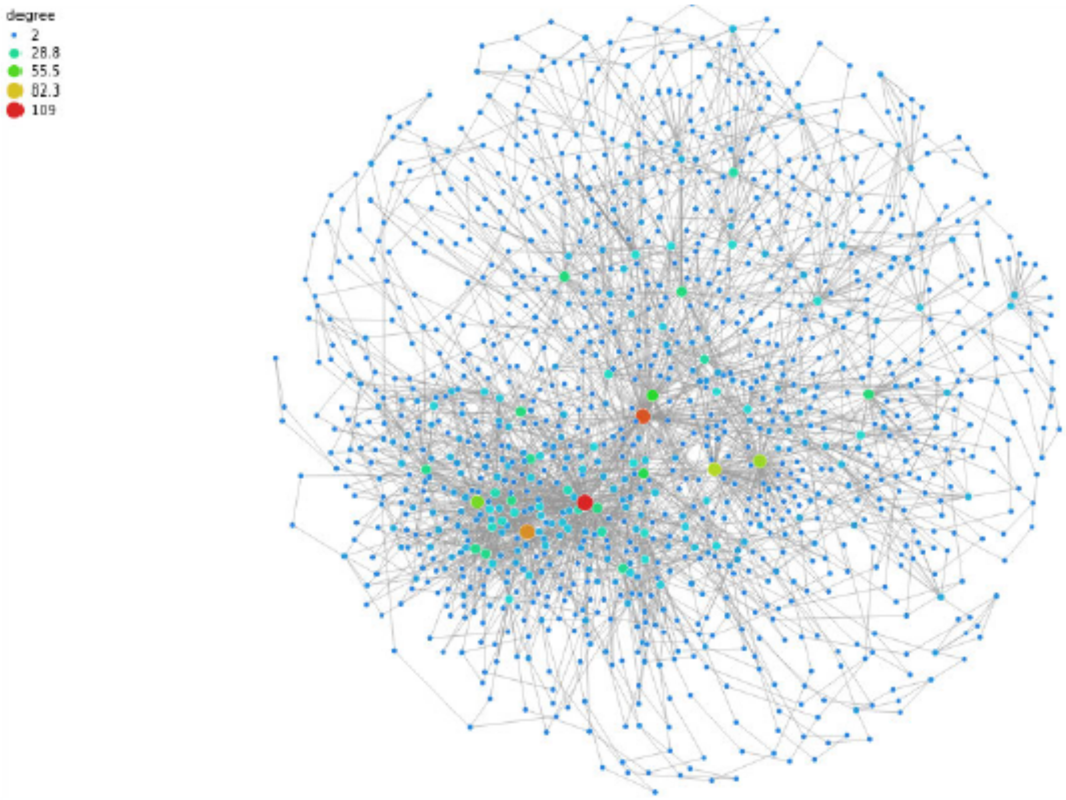


Figure 24: Network of retweets from the SDGs analysis.

In this network we can appreciate how the users tend to congregate around influential users, besides we can discern the presence of certain communities around these users.

In our study about eHealth and Citizen Science (entitled “An Analytics Approach to Health and Healthcare in Citizen Science Communications on Twitter”), which is currently under review in the eHealth SAGE journal, we decided to use the Louvain method to calculate communities. This method was changed in later analysis, since although this is a verified and widely used method, it does not compute overlaps in communities. This overlap may possibly explain how different nodes can belong to more than one community due to the fluctuating behaviour of the communities in social networks.

For this later analysis, we decided to use other algorithms and methods such as Link Partitioning. In Figure 25, we can see the result of the Louvain calculation.

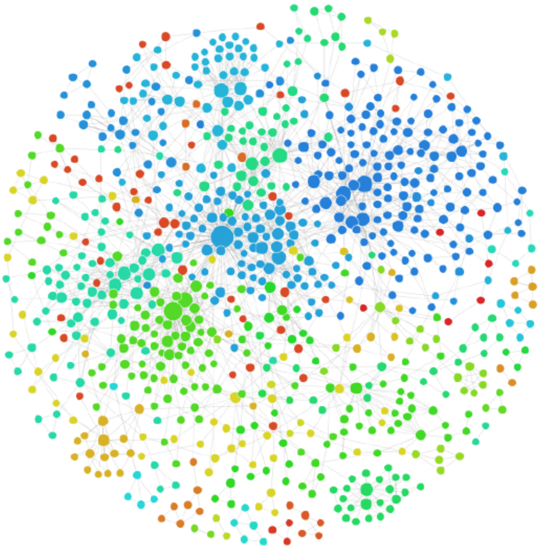


Figure 25: Second core of the graph of most retweeted users. Highlighting communities calculated with Louvain method

Besides studying the networks formed by the retweets, in our study about eHealth, we decided to also check the networks formed by mentioning users and the networks of hashtags. Retweets are not the only way to connect users, mentions are important since they can tell us who is active in creating boundaries with other members of the CS community. The idea behind analysing the hashtags in a network is that this way we would be able to visualise their distribution, thus understanding how the hashtags are used together and obtaining an initial view to the different topics that are linked to each other.

When reviewing the network of hashtags, in Figure 26, we see that they also tend to form communities according to the topic of the hashtag. The most used ones are placed in the outer parts of the network while the least used ones are in the middle, meaning that they are used alongside the others more frequently. To exemplify, we have the nodes (hashtags) of health, NCDs (Noncommunicable diseases, such as cancer or diabetes, also known as chronic diseases), SDGs, and wearables in the top part of the network. These are good examples of highly used hashtags alongside others, especially those around them. If we check those in the middle, we see hashtags like cancer and yoga or gerontology, not highly used but connected to all the hashtags in the network.

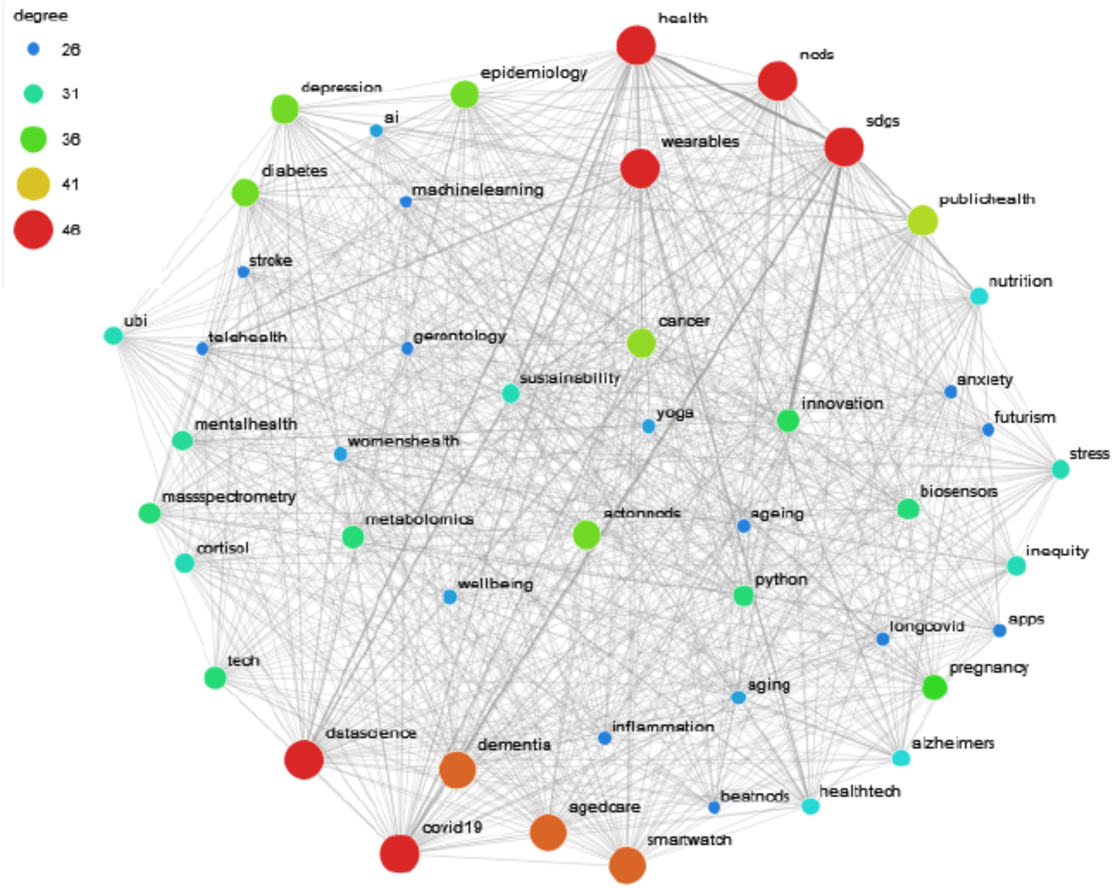


Figure 26: Network of hashtags

In the network of mentions we see that the users are far from each other, so there is a clear difference between who mentions who. One user, CitSciOZ, is highly mentioned so their content should be analysed to understand what makes them so influential. All this can be seen in Figure 27. Another interesting finding is that the accounts that are mentioned are projects or platforms in majority.

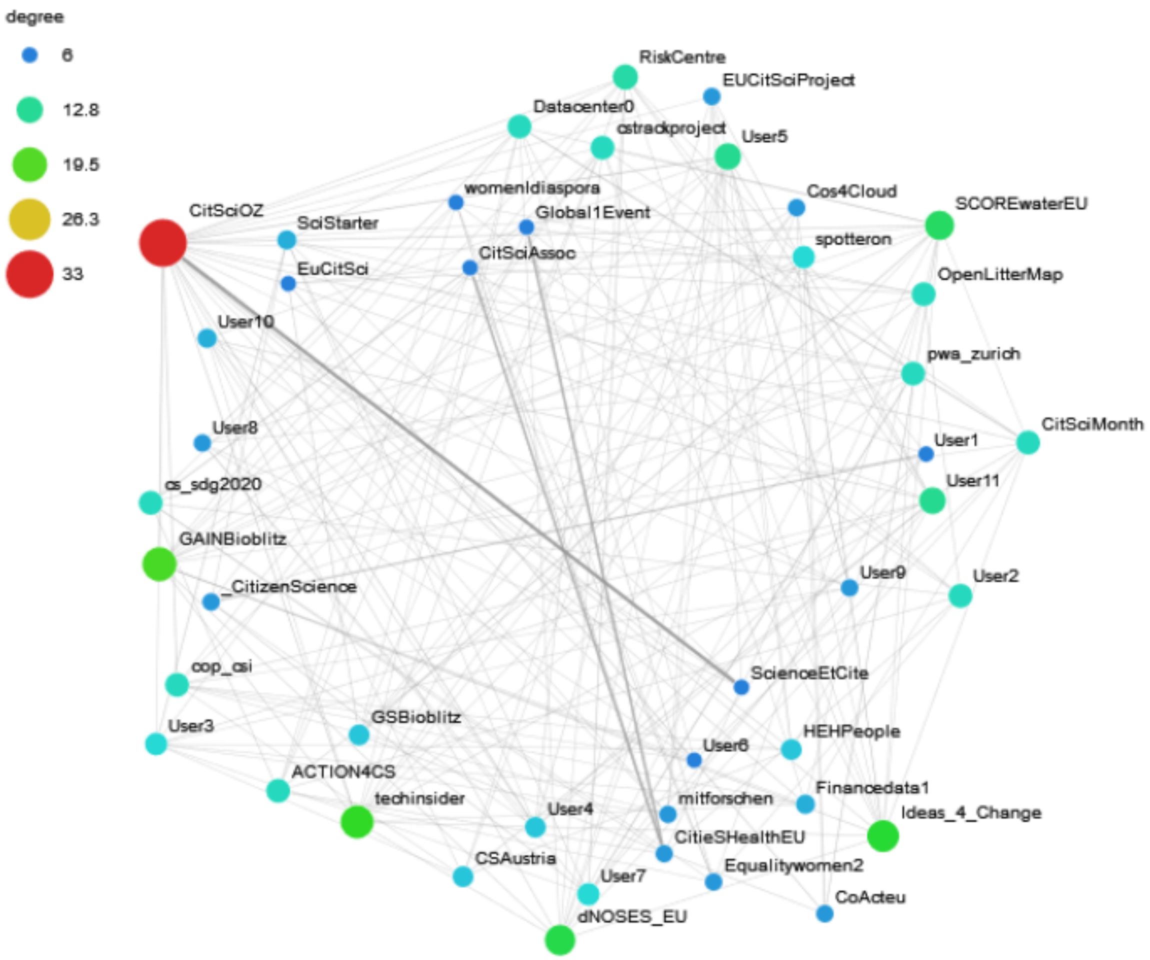


Figure 27: Network of mentions

More recently, we have improved the analysis of networks applying the information extracted using the Twitter API. This allows us to additionally consider the *follow* relation when constructing the network, as through the API, we can detect which users a particular user follows. For example, we can retrieve the followers list of user *u*, which is (*v, w, x*). To construct the network, we thus add (*u, v, w, x*) as nodes and accordingly, (*v, u*), (*w, u*) and (*x, u*) as edges to the network. However, if we follow this procedure and construct such a follow-network for our CS users which are part of the Lynguo data, the resulting network is extremely big with millions of edges, making any analyses computationally hard to realise. Thus, we explored ways to trim the network while maintaining the possibility to retrieve relevant insights. A common technique in this respect is the creation of *ego-networks* (Arnaboldi et al., 2016). To construct such networks, we only consider a pre-specified *ego* node and its neighbours, as well the connections between these neighbours. This "1.5-neighbourhood" is exemplified in Figure 28. As it can be seen, this leads to a significant reduction in network size.

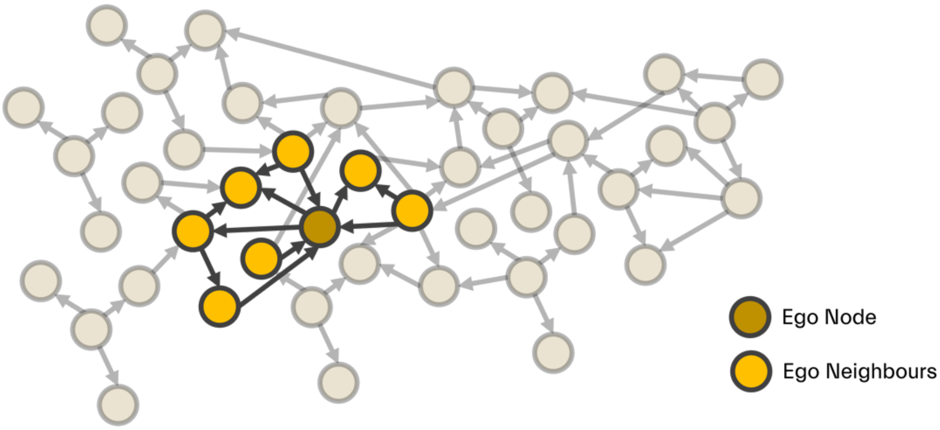


Figure 28: Example for the creation of an ego-network.

To follow this approach, we first defined relevant users to create these ego-networks. We chose two accounts, *EuCitSci*[7] (the official European Citizen Science Association account) and *SciStarter*[8] (the official account for the online CS platform) and extracted their ego-networks using the Twitter API. We chose these accounts because we wanted to specifically examine institutional accounts and how the community of private users and other institutions around them take form. For the extraction, we considered both the friends and the followers list of the associated neighbours, resulting in a complete ego-network based on the follow relation. By considering both directions for this extraction, we made sure that all neighbour nodes are in the resulting network, even if the follow connection between them and the ego is not mutual. Table 7 shows basic descriptive facts about these two accounts, such as their followers. As it can be seen, the community around *SciStarter* is substantially larger, which thus still results in an ego-network with 2 million edges.

Table 7: Information about the two accounts chosen for the ego-networks

Username	*@EuCitSci*	*@SciStarter*
Followers	5,310	14.3k
Friends	1,683	9,013
Likes received	4,088	49.8k
Nodes in ego-network:	5,868	17,698
Edges in ego-network:	179,794	2,004,840

In addition to the topological features which we can examine in the ego-networks, we included *CS-association* as another relevant node attribute. To calculate it, we checked whether the users in the networks appear in our Lynguo database, as this would mean they tweeted something containing CS keywords in the past two years. If they appear in this database, we assign them the “CS-associated” attribute, if not, they are labelled “non-CS-associated”. Accordingly, being or not being “CS-associated” does not necessarily imply a direct involvement or non-involvement in CS, yet it primarily indicates a clearly observable trace about an association with CS. In our analyses, we used this information to cluster our data into corresponding communities, in particular, a group that is CS-associated (CS-community) and one that is not (non-CS community).

[7] https://twitter.com/eucitsci

[8] https://twitter.com/SciStarter

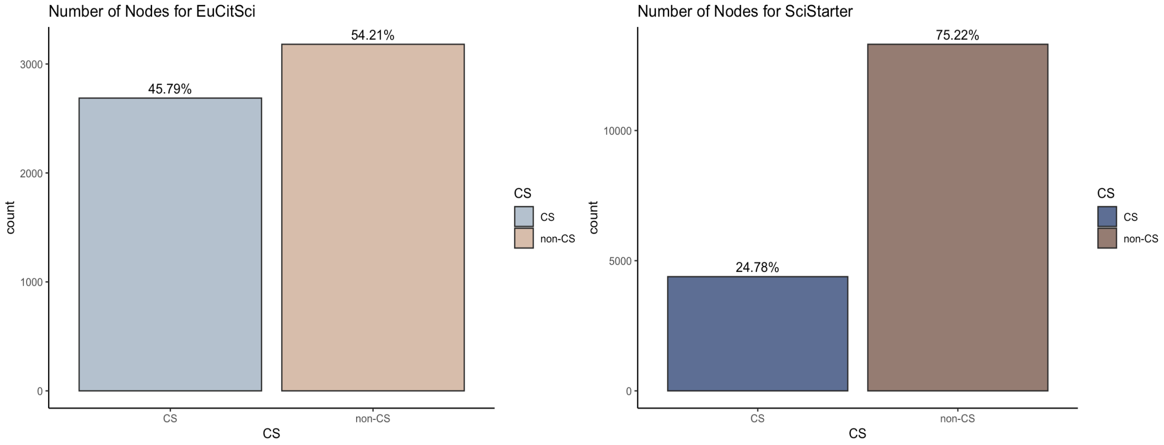


Figure 29: Distribution of CS-association across the two ego-networks

We first examined our two ego-networks with regard to this attribute. As it can be seen in Figure 29, there are significant differences between the two networks. While the users in *EuCitSci*'s ego-network are almost equally distributed regarding their CS-association, for *SciStarter*, we see that significantly more (75%) users are not associated with CS. However, these numbers represent all users appearing in the ego-network, irrespective of their relationship to the ego node (i.e., direction of follow-relationship).

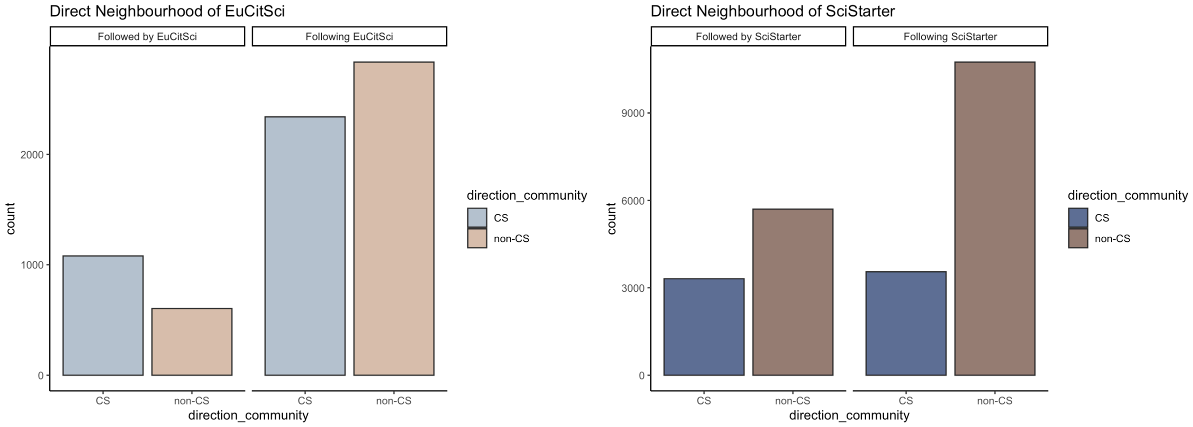


Figure 30: Distribution of followers/friends of EuCitSci and SciStarter

Thus, we also considered this, as can be seen in Figure 30: While more users who are not associated with CS follow *EuCitSci*, this is not the case for *EuCitSci*'s friends. There, we see that the majority of users are CS-associated. For *SciStarter*, this proportion is reversed, and most of the accounts that *SciStarter* follows are not CS-associated. The figure shows the total numbers. It is also relevant to mention that non-CS associated users represent a three-fold majority for the followers in *SciStarter*, as many people who are not actively tweeting or retweeting about citizen science (i.e., non-CS-associated) do still follow *SciStarter*. When we consider the follow-relationships on Twitter, we can derive conclusions about potential flows of information between users, because if user *u* follows user *v* (and therefore an edge *u, v* exists), information could potentially flow from user *v* to user *u*. Therefore, we can examine the relationships between different groups (e.g., CS-associations) and consider their prevalence and mutuality, as we can interpret these as a pre-requisite for information flow. We analysed this inter-community communication (see Figure 31) and observed differences with regard to how information could potentially flow.

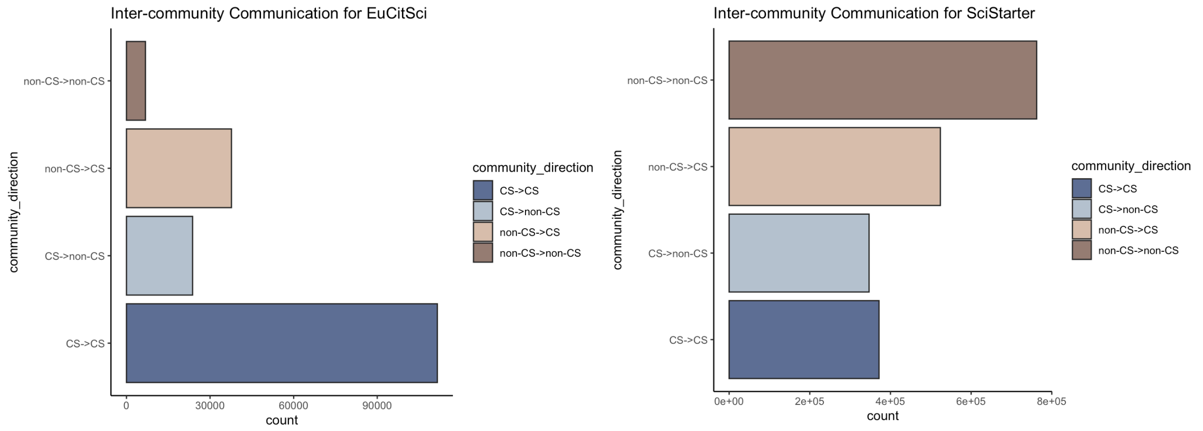


Figure 31: Inter-community communication between users with different CS-associations

For *EuCitSci*, most connections occur between members who are CS-associated, followed by non-CS to CS. Thus, information shared by users within this network potentially stays within the CS-associated community, and, to some extent, can diffuse to users who are not CS-associated. For *SciStarter*, this looks different, as such within-CS connections do not appear to be prevalent, yet most of the connections are from non-CS to other non-CS associated users. Thus, information which is shared in this network is more likely to stay outside CS-associated users. Interestingly, the second most prevalent type of connections are also non-CS to CS users, which further underscores that although there are differences between the two accounts, information from CS-associated users could still diffuse to non-associated users. Regarding such information flows, the mutuality of connections is of interest. As it can be seen in Figure 32, we counted the occurrence of mutual edges. Thus, for our CS-association communities, three different types are possible: mutual edges *within* the CS-associated community, *between* communities and *outside* the CS-association.

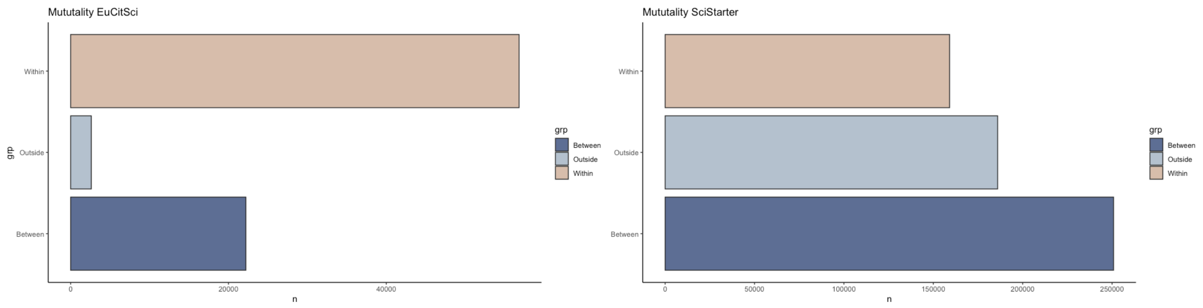


Figure 32: Mutuality of connections between different types of CS-associated users.

We find most mutual edges *within* the CS community for *EuCitSci*, while mutual edges *between* different communities appear to be highly prevalent for *SciStarter.* Thus, mutual information exchange most likely occurs between CS-associated people in the ego-network for *EuCitSci*, while for *SciStarter*, between users of different CS-associations.

The described analyses show that the creation of ego-networks in combination with certain community-related attributes can provide relevant insights, while still decreasing the information load and complexity by only considering a specific proportion of the whole Twitter-sphere. Doing so can reveal the prevalence of potential paths for information to flow between actors. Accordingly, this can help researchers to improve their outreach by specifically targeting their tweets to an audience that is not necessarily associated with citizen science (e.g., *SciStarter*), or trying to increase their reach within such audiences (*e.g., EuCitSci*). An extension to these findings would be the inclusion of actual tweet data like retweets to the existing follow-relational structure, as it would allow for a more explicit analysis of actual information flow as opposed to potential information flow. Future perspectives of ego-networks formed by similar profiles to EUCItsci and SciStarter could be included as well as the idea of incorporating actual retweets into these ego-networks to the model's information.

3.3.4. Triangulation of Twitter data between WPs

Inside the project, other work packages have been working with data from different sources. WP2 had the data from the projects extracted from the repositories of Citizen Science and WP4 had data from the surveys they have been conducting. One of the main aspects they have worked in is the perceived learning gains of the participants in Citizen Science projects vs. the desired or intended learning outcomes from the project initiators. The goal of the triangulation is to align the results from the different sources and have insight on the expectations from the projects and the actual comments of the individuals when discussing the learning outcomes. WP2 has been studying mainly the expected learning gains by means of the description, where the project initiators provide some ideas of what will be learned by the individuals who participate. WP4 started a survey destined to obtain comments from people who have been involved in a CS project about what they learnt once they finished their collaboration.

To compare the results from the different databases, we decided to check Twitter for relevant information. To analyse this, we used the categories proposed in Phillips 'publication (Phillips et al., 2018) but only those that were specific enough for our purpose:

- Content, Process and Nature of Science Knowledge
- Skills of Science Inquiry
- Using Technology
- Training and didactic materials provided by the project
- Access to knowledge and data produced by the project

The selection of these categories was done because we slightly modified their coding scheme to adapt it to our text corpus. WP2 then tested out the keywords extracted from the qualitative content analysis using nCoder and BERT, and found that for many categories, the keywords were not specific enough, so we ended up with the following set of keywords for each category:

Table 8: Set of keywords

Attribute	Keywords
Content, Process and Nature of Science Knowledge	find out, learn about, learn to, learn why, learn more about, learning, discover, guide, get to know, become familiar with, deepen their skills, deepen their knowledge, increase knowledge, use a range of methods, understand
Skills of Science Inquiry	upload, tell us, enter, insert, report, answer, fill out, fill in, send, transcribe, collect, tag, find, identify, locate, search, take a photo, note, record, observe, make a video, look out for, map, measure, sensor, gather, monitor, extract, count, take measurements, observations,

	transcribe, co-research, come up with, analyse, analyse, mark, determine, interpret, discuss, transmit, write, present, deliver, proposal, suggest, policy making, conclusions, recommendations
Using Technology	platform, online platform, web platform, app, video, microscope, tool, mobile device, device, GPS, recordings, application, upload, photo, digitise, digitalise, sensor, device, online, interactive, download, website, Android, iOS
Training and didactic materials provided by the project	guide, instructions, how to, recommendations, tips, educational material, feedback, lesson, workshop, educational, advice, booklet, training session, coach, tutorial, webinar
Access to knowledge and data produced by the project	access, publicly available, made available, open code, open data, open access

With the keywords ready for each category, we split the Lynguo dataset into tweets written by project, platforms, and users. For differentiation, we selected 11 projects in which we find Spanish, English, Finnish and German projects to fit all the different groups involved in triangulation. Two of these projects had their own Twitter account and the rest were managed by platforms (organisations that tweet about different projects). With this separation, besides selecting only those tweets about the projects of our interest using once more our own filter by keywords function, we had the datasets for projects and for platforms. To select the tweets from the users, we removed from the dataset every tweet belonging to a project or a platform using a list of CS projects previously found in the Twitter data. For each group, we analysed the distribution of the categories, once more using the filter by keywords function with the different keywords from the five categories. Then, we applied some of the methods described in the previous sections. The main techniques used were: filtering by keywords to extract the tweets from the categories, a hashtag analysis, an extraction of the most used words and another new method which is the analysis of impact. This impact is calculated for each tweet according to the number of followers of the account. This is a value provided by Lynguo and ranges from 0 to 100.

The results show a gap between the project coordinators’ perspective on educational aspects of their projects and the way citizen scientists perceived learning in these projects. The words used by the users in the different categories, and the distribution of categories (meaning which categories are more discussed) in Twitter do not align with the results from the survey and the results from the analysis of the project´s description. These findings suggest that the participants in CS projects could access online training and online material as it is stated in the project´s descriptions. However, the results from the analysis of the survey and the Twitter data suggest that the learning gains of users are not always aligned to the expectancies of the projects. In other words, users learned something different during their participation to what the project´s description said they would learn. Analysing the complete panorama of CS and learning, we believe that communication is important in CS to promote participation, but the expectations from coordinators should be aligned with users’, i.e., participants’, expectations.

Section 4: Interfacing and exportation of results

Task 3.4 ("Interfacing and exportation of results") has been conceived as an interaction between the science communication platform (WP5 led by URJC) and the data-generating activities in WP3. The platform has now essentially taken the form of the eMagazine, which uses standardised formats and mechanisms for rendering content. The original idea was to have semi-automatic chains of information processing from the web analytics to the platform. Based on the agreed-upon standards, a "manual" step for feeding results into the eMagazine was needed. The analytics tool chain provided visualisation that were easily transferable into "graphical articles".

Another channel of exportation and distribution of results (including tools and applications) have been workshops and webinars. The first in the newly established series of CS Track webinars was about the AWB, which had already been revealed in three other interactive demo events before (one internal workshop and two external demos). Together with the work from the participation/motivation case study, it was finally demonstrated in the analytics workshop at the CS Track Symposium following the ECSA Conference 2022 in Berlin.

This section includes a subsection that lists the output of WP3 in terms of "normal" scientific publications, here restricted to peer-reviewed papers at international conferences and journal articles.

As part of CS Track's open data and open-source strategy, data and code sources as well as technical documents are being uploaded to a Zenodo archive managed by CS Track. Code archives are also available under GitHub or GitLab (final release pending).

The results of analytics studies were fed into the discussions underlying the policy recommendations. These recommendations are typically based on multiple sources of evidence and may include normative aspects. Interestingly, it was possible to derive some policy recommendations from specific analytics results even before combining these with findings from surveys and interviews (subjective data). The last subsection contains such recommendations.

4.1. Publication and distribution of results through the eMagazine

The eMagazine is CS Track's main channel for the communication and distribution of results to a wider interested public. As explained above, web analytics results have mostly used the format of graphical articles, allowing for exporting previously generated data visualisations with shorter textual comments.

- What are the predominant research areas in citizen science projects? (RIAS, March 2021)
- Are citizen science projects multidisciplinary research activities? (RIAS, March 2021)
- Citizen Science and open learning: A Twitter perspective (URJC, July 2021)
- How social network analysis reveals significant variables in Citizen Science projects: The Chimp & See case (RIAS, November 2022)
- How do different participants contribute to the knowledge-building discourse in online citizen science projects? (UPF, February 2022)

- A short introduction to the CS Track Analytics Workbench (RIAS, May 2022)
- SDG discussion in the Citizen Science community of Twitter (URJC, June 2022)
- The importance of the few – how a minority of power users shape most of the discourse in CS forums (June 2022, RIAS)
- How are tweets distributed for each Sustainable Development Goal within the Citizen Science Twitter community? (URJC, July 2022)
- How does the Citizen Science community use hashtags when discussing eHealth? (URJC, July 2022)

4.2. Propagation of methods, tools and results through workshops and webinars

The Analytics Workbench (AWB) has been in the centre of one online workshop in November 2021 and of the first CS Track Webinar in March 2022. The AWB was also a main topic of the presentation given in January 2022 by Cleo Schulten and Ulrich Hoppe (RIAS) in the "Digital Citizen Science" speaker series of the Karlsruhe Institute of Technology (KIT). It was also included in the program of the “Engaging Citizen Science Conference” in April 2022 in Aarhus (DK) as an interactive demo. The objectives of these demo activities included the propagation of ideas, tools and methods to specific interest groups (especially CS researchers and project managers) as well as the provision of user feedback and evaluation data for system improvement (see section 2.2). Both the workshop and the webinar were limited in participation due to the structure of hands-on activities with pre-selected project samples. In both cases, the limit of 30 registrants was reached with a rich international participation and diversity of stakeholders. Figure 33 shows the distribution of countries for the participants of the webinar.

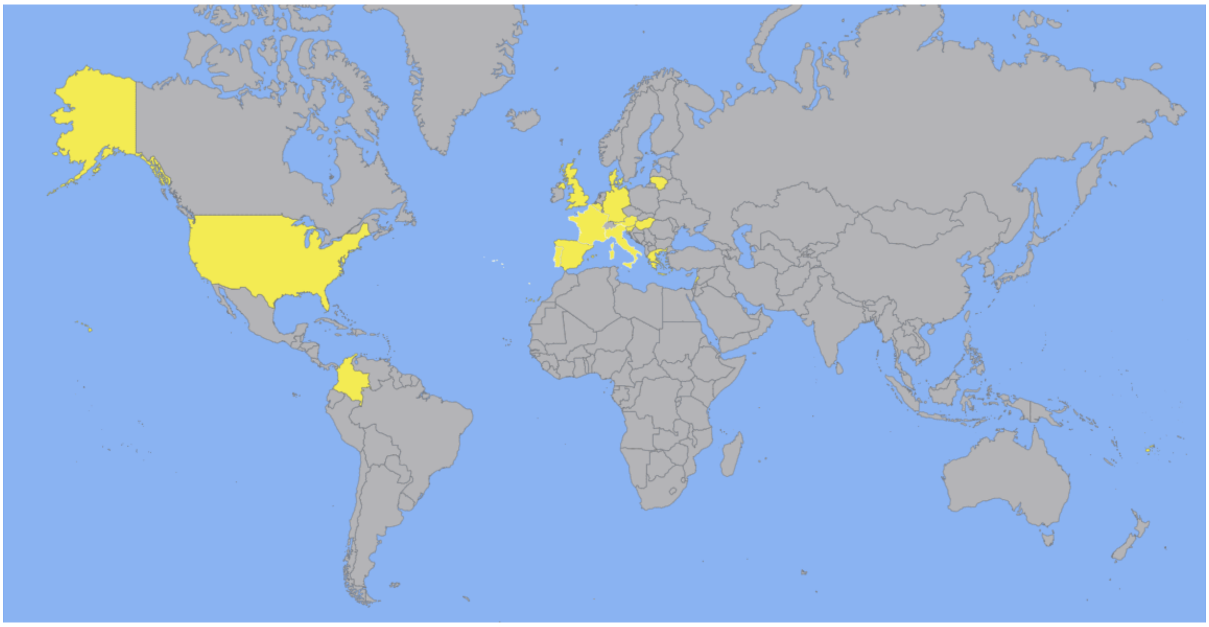


Figure 33: International distribution of participants of the second AWB workshop (webinar)

For the second webinar in the CS Track series ("Describing your citizen science project – How to present your project and recruit volunteers"), a simplified version of the AWB ("AWB lite") was created that allowed for entering a project description (i.e., a text) followed by the generation of RA and SDG associations. In this way, the generative functions of the AWB were provided without giving free access to project database, which is still problematic because of personal information details.

The CS Track Symposium at the ECSA conference in Berlin (November 2022) featured another workshop dedicated to analytics methods and tools. Here again, the AWB was interactively presented followed by an introduction to participation analysis. For this analysis, an interactive presentation of results in the form of a Quarto notebook with hands-on generation of network visualisations, results and user trajectories was used to demonstrate how the developed analysis techniques can be used to track the trajectories of individual participants and to examine the discussion structure within CS forums.

This notebook was already described in section 3.2.5 and can be found in Annex 6.2 or by visiting https://t1p.de/ecsa-symposium (link will be valid until March 31st, 2023).

4.3. Peer-reviewed international publications

As of November 2022, the following peer-reviewed papers or articles originating (mainly) from WP3 have appeared in or have been accepted for the following venues:

[1] Amarasinghe, I., Manske, S., Hoppe, H. U., Santos, P., & Hernández-Leo, D. (2021). Using network analysis to characterize participation and interaction in a citizen science online community. In *International Conference on Collaboration Technologies and Social Computing* (CollabTech 2021) (pp. 67-82). Springer LNCS 12856, Cham.

[2] 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 2021) (pp. 6-8). IEEE Press.

[3] Roldán-Álvarez, D., Martínez-Martínez, F., Martín, E., & Haya, P. A. (2021). Understanding discussions of Citizen Science around Sustainable Development Goals in Twitter. *IEEE Access*, 9, 144106-144120.

[4] Hoppe, H. U., Schulten, C., Santos, P., Calvera, M., DeGroot, R., & Golumbic, Y. (2022). Between exoplanets and planetary health: Viewing Citizen Science through the SDG lens. Proceedings of the *Conference of the European Citizen Science Association* (ECSA 2022). Berlin, October 2022. https://2022.ecsa-conference.eu/files/ecsa/Bilder/ECSA2022_Conference_Proceedings.pdf

[5] Krukowski, S., Amarasinghe, I., Gutiérrez-Páez, N. F., & Hoppe, H. U. (2022). Does volunteer engagement pay off? An analysis of user participation in online Citizen Science projects. In *International Conference on Collaboration Technologies and Social Computing* (CollabTech 2022) (pp. 67-82). Springer LNCS 13632, Cham.

[6] De-Groot, R., Golumbic, Y. N., Martínez Martínez, F., Hoppe, H. U., & Reynolds, S. (2022). Developing a framework for investigating Citizen Science through a combination of web analytics and social science methods - the CS Track perspective. *Frontiers in Research Metrics and Analytics*, 7. https://www.frontiersin.org/articles/10.3389/frma.2022.988544

[7] Martínez-Martínez, F., Martín, E., Roldán-Álvarez, D., & Hoppe, H. U. (2023). An analytics approach to health and healthcare in Citizen Science communications on Twitter. *Digital Health* (SAGE), to appear in vol. 9 (accepted Nov. 28, 2022).

When looking at this list of scientific publications, it is important to consider that for most contributors their primary disciplinary context is computer science. This is clearly visible in the choice of the venues. Especially the conferences (*ICALT*, *CollabTech*) reflect links based on the existing and established research orientations of the participating groups with their grown communities. In both cases, these venues have long-established peer-reviewing committees and edited books as publications (*IEEE Press* and *Springer LNCS*). The contribution to the *ECSA conference* has also undergone a peer review, yet the publication as such is more informal. We still see it as important since it connects the work to the European and international CS community.

Whereas the journal *IEEE Access* again reflects the computer science orientation, *Frontiers in Research Metrics and Analytics* represents a meta-level methodological perspective and finally *Digital Health* (SAGE) stands for a more domain-specific application perspective.

4.4. Sharing of source code and datasets (GitHub/GitLab and Zenodo)

As part of the Open Data strategy, CS Track uses Zenodo and other platforms for giving public access to research results and techniques in the form of documents and datasets. To this end, datasets and analysis scripts/notebooks were primarily made available through Zenodo, while source code was made available through GitHub/GitLab.

4.4.1. Source Code

The source code used in the described tools and analyses was made available through a collaborative shared GitHub account (i.e., organisation). In this organisation, several repositories containing the associated open-source code can be found. These can be accessed by visiting the following link:

https://github.com/CS-Track-Code

For this deliverable, certain repositories are of special relevance, the following in particular:

- **AWB**
    - Code associated with the frontend, middleware and backend of the Analytics Work-bench (see section 2)
    - https://github.com/CS-Track-Code/analytics-workbench
- **Participation and Motivation**
    - Code that was used to generate the data and extract networks within the context of the Zooniverse participation analysis (see section 3.2)
    - https://github.com/CS-Track-Code/zooniverse-network-extraction

All repositories contain a readme file with further explanations and instructions on how to get started.

4.4.2. Datasets and analysis tools

The data overlap matrices underlying the study of **interdependencies of research areas (RAs) and SDGs** (section 3.1.2) have been uploaded to Zenodo in two versions: One including all RAs that were at least assigned to 10 projects in the sample and another one with broader set of RAs based on minimally 5 occurrences. The data shown in the table are based on the more version (min = 10). The tables can be found here on Zenodo:

https://doi.org/10.5281/zenodo.7353663

The data used for our analyses on **participation and motivation** in Zooniverse forums was uploaded to Zenodo in the form of multiple distinct (dataset) publications. Specifically, the uploads represent the three parts necessary to potentially reproduce and extend the findings, and additional uploads:

Raw data

The raw data which represents the basis for the network extraction and analysis steps was uploaded to Zenodo and is constituted of several .json files. Each data point (i.e., comment) was anonymised to replace the original usernames by an alphanumerical string. More information can be found in Zenodo entry:

https://doi.org/10.5281/zenodo.7357835

Result tables, network files and corresponding analysis document

The data basis for our analyses (node/edge lists, result tables) were also uploaded to Zenodo in a structured way, primarily as .csv files. A corresponding Quarto-notebook was uploaded which can be used to replicate the results described in this deliverable. More information (specifically on the data structure) can be found in Zenodo entry:

https://doi.org/10.5281/zenodo.7357746

Network files

To easily visualise or interpret the data with common network analysis applications (e.g., Gephi, Pajek), we exported the extracted networks for each of the analysed projects as .gml and .gexf files. They can be found here:

https://doi.org/10.5281/zenodo.7356425

For our **Twitter analyses**, we also made available the datasets and associated analysis documents. For the analyses on SDGs, eHealth and related aspects, these can be found here:

https://doi.org/10.5281/zenodo.7373532

For the data that is the basis of the ego-network analyses, all corresponding data can be found here:

https://doi.org/10.5281/zenodo.7372308

Further, the specific result tables (and steps to reproduce) can be found in another Zenodo upload that was made:

https://doi.org/10.5281/zenodo.7360192

4.5. Deriving specific policy recommendations from analytics results

Usually, the results of computational analytics have been integrated with findings from surveys and interviews in a triangulation process before drawing conclusions in terms of policy recommendations and statements assessing the societal impact and contribution of CS. In our discussions, however, we have seen that certain findings based on analytics techniques could be directly fed into specific policy statements:

- Monitoring tools (see section 3.2) can identify particularly engaged volunteers as a basis for providing positive feedback and possible promotions to roles with higher responsibility.
- Content analysis of research areas and SDGs indicates a caveat related to using the resonance of CS activities with SDGs carefully: Be aware that traditional areas of CS such astronomy/astrophysics might be left behind without realising.
- The analysis of research areas yields a distinction between core disciplines of science and instrumental areas of support (e.g., climatology vs. remote sensing): Decision makers should be aware that instrumental areas are also important, and CS projects may generate valuable advances also in these instrumental areas (such as "remote sensing").
- Content analysis techniques can used to identify skills (STEM and soft skills) likely to be promoted through participation in CS projects. This is a possible information source for accreditation.
- Supporting citizen participation in science as a means to increase rational dispute rather than emotionally loaded discussions.

Section 5: References

Aggarwal, N., Asooja, K., Bordea, G., & Buitelaar, P. (2015). Non-orthogonal explicit semantic analysis. *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, 92–100. https://doi.org/10.18653/v1/S15-1010

Amarasinghe, I., Manske, S., Hoppe, H. U., Santos, P., & Hernández-Leo, D. (2021). Using network analysis to characterize participation and interaction in a citizen science online community. In D. Hernández-Leo, R. Hishiyama, G. Zurita, B. Weyers, A. Nolte, & H. Ogata (Eds.), *Collaboration Technologies and Social Computing* (pp. 67–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-85071-5_5

Arnaboldi, V., Conti, M., La Gala, M., Passarella, A., & Pezzoni, F. (2016). Ego network structure in online social networks and its impact on information diffusion. *Computer Communications*, *76*, 26–41. https://doi.org/10.1016/j.comcom.2015.09.028

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, *3*(null), 993–1022.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., & Wilderman, C. C. (2009). Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. In *Online Submission*. https://eric.ed.gov/?id=ED519688

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

De-Groot, R., Golumbic, Y. N., Martínez Martínez, F., Hoppe, H. U., & Reynolds, S. (2022). Developing a framework for investigating citizen science through a combination of web analytics and social science methods—The CS Track perspective. *Frontiers in Research Metrics and Analytics*, *7*. https://www.frontiersin.org/articles/10.3389/frma.2022.988544

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

Fritz, S., See, L., Carlson, T., Haklay, M. (Muki), Oliver, J. L., Fraisl, D., Mondardini, R., Brocklehurst, M., Shanley, L. A., Schade, S., Wehn, U., Abrate, T., Anstee, J., Arnold, S., Billot, M., Campbell, J., Espey, J., Gold, M., Hager, G., … West, S. (2019). Citizen science and the United Nations Sustainable Development Goals. *Nature Sustainability*, *2*(10), Article 10. https://doi.org/10.1038/s41893-019-0390-3

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, 1606–1611.

Haralambous, Y., & Klyuev, V. (2014). *Thematically reinforced explicit semantic analysis* (arXiv:1405.4364). arXiv. https://doi.org/10.48550/arXiv.1405.4364

Hecking, T., Chounta, I. A., & Hoppe, H. U. (2017). Role Modelling in MOOC Discussion Forums. *Journal of Learning Analytics*, *4*(1), Article 1. https://doi.org/10.18608/jla.2017.41.6

Hoppe, H. U., Schulten, C., Santos, P., Calvera, M., DeGroot, R., & Golumbic, Y. (2022). Between exoplanets and planetary health: Viewing Citizen Science through the SDG lens. *Proceedings of ECSA 2022*. ECSA 2022, Berlin. https://2022.ecsa-conference.eu/files/ecsa/Bilder/ECSA2022_Conference_Proceedings.pdf

Krukowski, S., Amarasinghe, I., Gutiérrez-Páez, N. F., & Hoppe, H. U. (2022). Does volunteer engagement pay off? An analysis of user participation in online Citizen Science projects. In L.-H. Wong, Y. Hayashi, C. A. Collazos, C. Alvarez, G. Zurita, & N. Baloian (Eds.), *Collaboration Technologies and Social Computing* (pp. 67–82). Springer International Publishing. https://doi.org/10.1007/978-3-031-20218-6_5

Martínez Martínez, F., Roldán-Álvarez, D., & Martín, E. (2022a). *SDG discussion in the Citizen Science community of Twitter*. https://cstrack.eu/format/graphical-article/sdg-discussion-in-the-citizen-science-community-of-twitter/

Martínez Martínez, F., Roldán-Álvarez, D., & Martín, E. (2022b, July 8). *How are tweets distributed for each Sustainable Development Goal within the Citizen Science Twitter community?* https://cstrack.eu/format/graphical-article/how-are-tweets-distributed-for-each-sustainable-development-goal-within-the-citizen-science-twitter-community/

Martínez Martínez, F., Roldán-Álvarez, D., & Martín, E. (2022c, August 14). *How does the Citizen Science community use hashtags when discussing e-Health?* https://cstrack.eu/format/graphical-article/how-does-the-citizen-science-community-use-hashtags-when-discussing-e-health/

Mazumdar, S., & Thakker, D. (2020). Citizen Science on Twitter: Using Data Analytics to Understand Conversations and Networks. *Future Internet*, *12*(12), 1–22.

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. http://arxiv.org/abs/1802.03426

Michalak, K. (2015). Online Localization of “Zooniverse” Citizen Science Projects—On the Use of Translation Platforms as Tools for Translator Education. *Teaching English with Technology*, *15*(3), 61–70.

Moczek, N., Voigt-Heucke, S. L., Mortega, K. G., Fabó Cartas, C., & Knobloch, J. (2021). A self-assessment of European Citizen Science projects on their contribution to the UN Sustainable Development Goals (SDGs). *Sustainability*, *13*(4), Article 4. https://doi.org/10.3390/su13041774

Phillips, T., Porticella, N., Constas, M., & Bonney, R. (2018). A framework for articulating and measuring individual learning outcomes from participation in Citizen Science. *Citizen Science: Theory and Practice*, *3*(2), Article 2. https://doi.org/10.5334/cstp.126

Ponciano, L., & Brasileiro, F. (2014). Finding Volunteers’ Engagement Profiles in Human Computation for Citizen Science Projects. *Human Computation*, *1*(2). https://doi.org/10.15346/hc.v1i2.12

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Rohden, F., Kullenberg, C., Hagen, N., & Kasperowski, D. (2019). Tagging, Pinging and Linking – User Roles in Virtual Citizen Science Forums. *Citizen Science: Theory and Practice*, *4*(1), Article 1. https://doi.org/10.5334/cstp.181

Roldán-Álvarez, D., & Martínez Martínez, F. (2021). Citizen Science and open learning: A Twitter perspective. *E-Magazine*. https://cstrack.eu/format/graphical-article/citizen-science-and-open-learning-a-twitter-perspective/

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

Roldán-Álvarez, D., Martínez-Martínez, F., Martín, E., & Haya, P. A. (2021). Understanding discussions of Citizen Science around Sustainable Development Goals in Twitter. *IEEE Access*, *9*, 144106–144120. https://doi.org/10.1109/ACCESS.2021.3122086

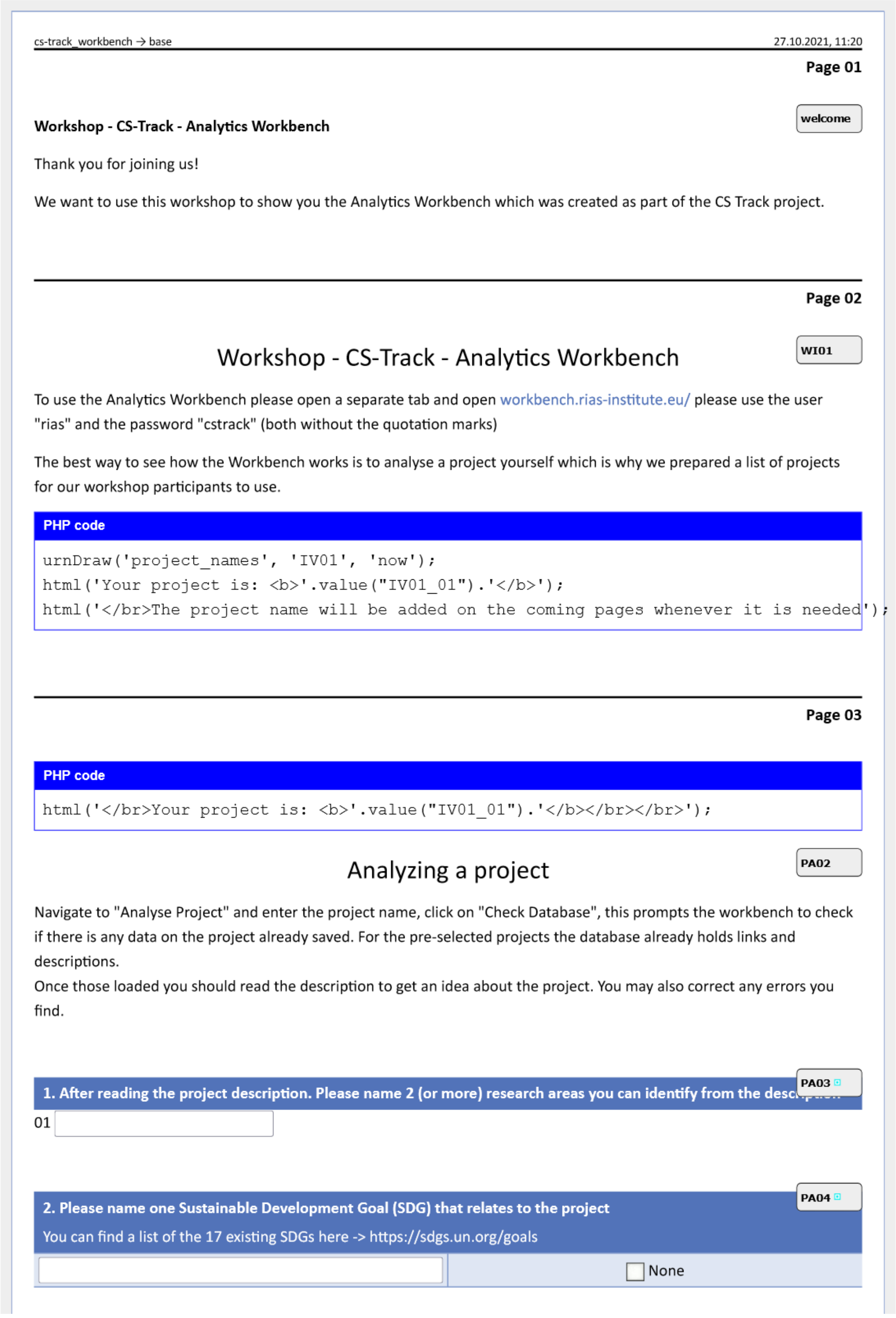
Sarirete, A., & Brahimi, T. (2014). Enabling communities of practice within MOOCs. *2014 International Conference on Web and Open Access to Learning (ICWOAL)*, 1–4. https://doi.org/10.1109/ICWOAL.2014.7009232

Scholl, P., Böhnstedt, D., Domínguez García, R., Rensing, C., & Steinmetz, R. (2010). Extended explicit semantic analysis for calculating semantic relatedness of web resources. In M. Wolpers, P. A. Kirschner, M. Scheffel, S. Lindstaedt, & V. Dimitrova (Eds.), *Sustaining TEL: From Innovation to Learning and Practice* (pp. 324–339). Springer. https://doi.org/10.1007/978-3-642-16020-2_22

Schrepp, M., Thomaschewski, J., & Hinderks, A. (2017). Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(Regular Issue). https://www.ijimai.org/journal/bibcite/reference/2634

Section 6: Annex

6.1. AWB Evaluation Questionnaire



cs-track_workbench → base 27.10.2021, 11:20

Page 01

welcome

Workshop - CS-Track - Analytics Workbench

Thank you for joining us!

We want to use this workshop to show you the Analytics Workbench which was created as part of the CS Track project.

Page 02

WI01

Workshop - CS-Track - Analytics Workbench

To use the Analytics Workbench please open a separate tab and open workbench.rias-institute.eu/ please use the user "rias" and the password "cstrack" (both without the quotation marks)

The best way to see how the Workbench works is to analyse a project yourself which is why we prepared a list of projects for our workshop participants to use.

PHP code

```
urnDraw('project_names', 'IV01', 'now');
html('Your project is: <b>'.value("IV01_01").'</b>');
html('</br>The project name will be added on the coming pages whenever it is needed');
```

Page 03

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b></br></br>');
```

PA02

Analyzing a project

Navigate to "Analyse Project" and enter the project name, click on "Check Database", this prompts the workbench to check if there is any data on the project already saved. For the pre-selected projects the database already holds links and descriptions.
Once those loaded you should read the description to get an idea about the project. You may also correct any errors you find.

PA03

1. After reading the project description. Please name 2 (or more) research areas you can identify from the desc

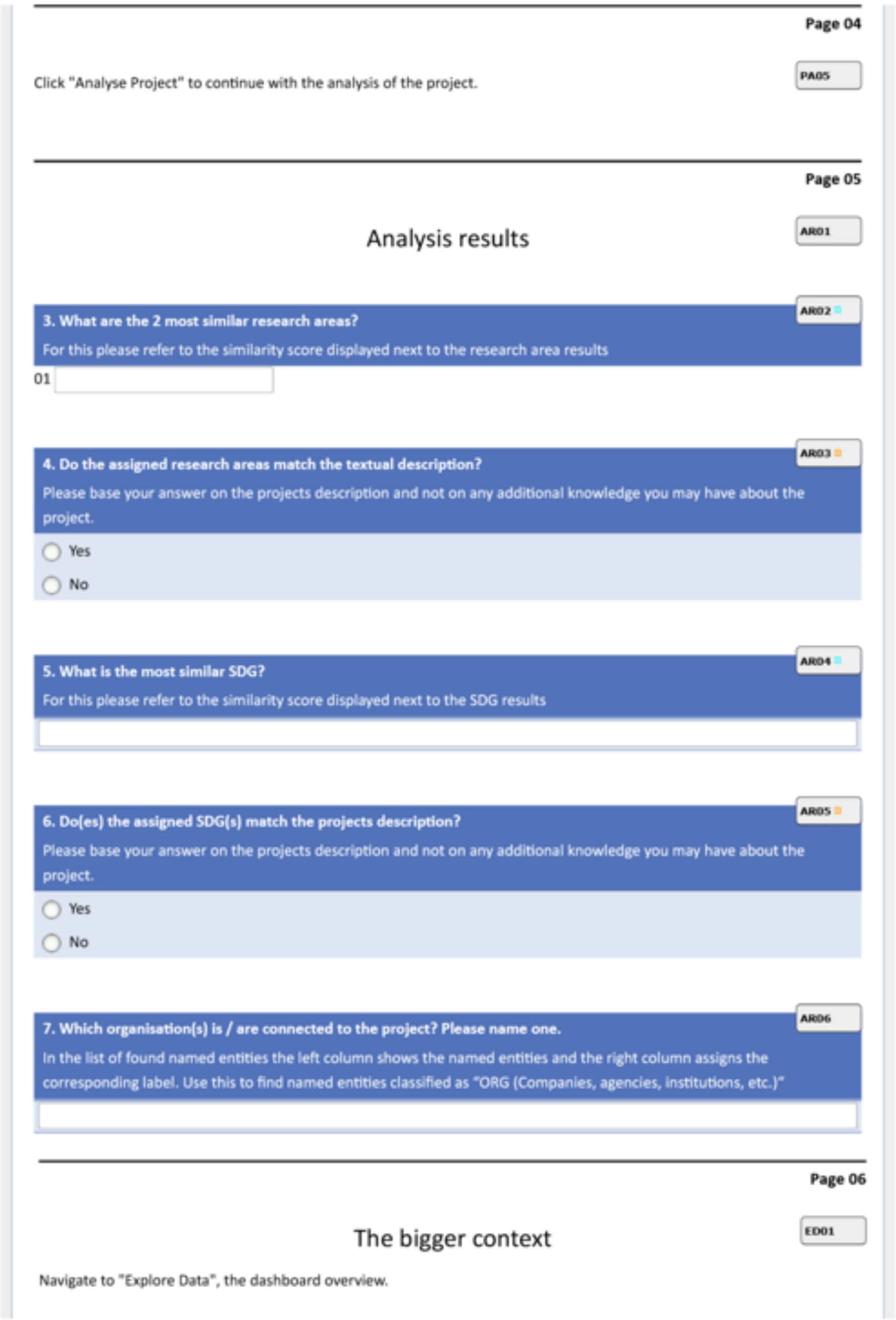
01

PA04

2. Please name one Sustainable Development Goal (SDG) that relates to the project

You can find a list of the 17 existing SDGs here -> https://sdgs.un.org/goals

□ None



Page 04

Click "Analyse Project" to continue with the analysis of the project.

Page 05

Analysis results

3. What are the 2 most similar research areas?

For this please refer to the similarity score displayed next to the research area results

01

4. Do the assigned research areas match the textual description?

Please base your answer on the projects description and not on any additional knowledge you may have about the project.

○ Yes

○ No

5. What is the most similar SDG?

For this please refer to the similarity score displayed next to the SDG results

6. Do(es) the assigned SDG(s) match the projects description?

Please base your answer on the projects description and not on any additional knowledge you may have about the project.

○ Yes

○ No

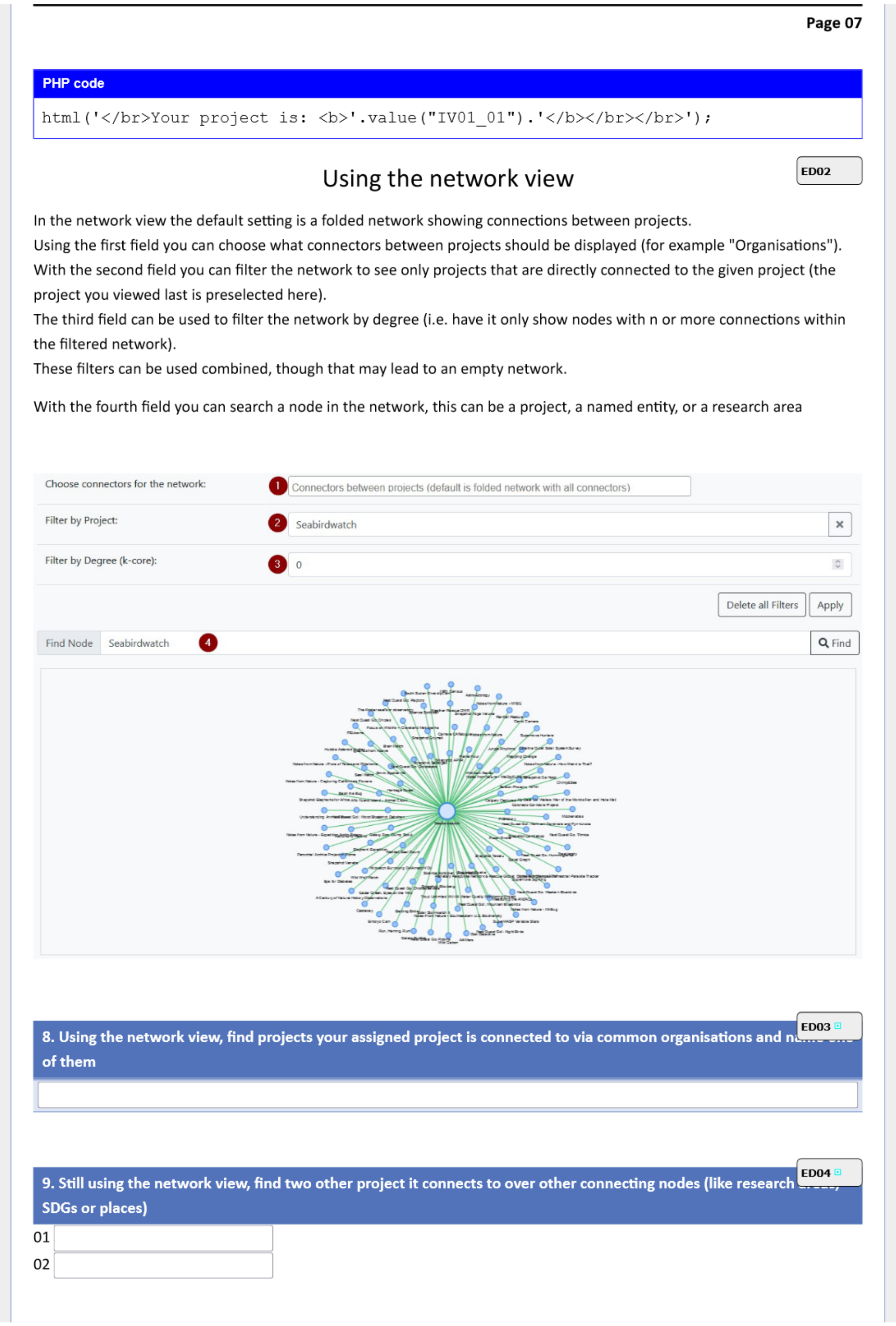
7. Which organisation(s) is / are connected to the project? Please name one.

In the list of found named entities the left column shows the named entities and the right column assigns the corresponding label. Use this to find named entities classified as "ORG (Companies, agencies, institutions, etc.)"

Page 06

The bigger context

Navigate to "Explore Data", the dashboard overview.



Page 07

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b></br></br>');
```

Using the network view

ED02

In the network view the default setting is a folded network showing connections between projects.
Using the first field you can choose what connectors between projects should be displayed (for example "Organisations").
With the second field you can filter the network to see only projects that are directly connected to the given project (the project you viewed last is preselected here).
The third field can be used to filter the network by degree (i.e. have it only show nodes with n or more connections within the filtered network).
These filters can be used combined, though that may lead to an empty network.

With the fourth field you can search a node in the network, this can be a project, a named entity, or a research area

Choose connectors for the network:	1	Connectors between projects (default is folded network with all connectors)
Filter by Project:	2	Seabirdwatch
Filter by Degree (k-core):	3	0

Find Node: Seabirdwatch 4

8. Using the network view, find projects your assigned project is connected to via common organisations and name one of them

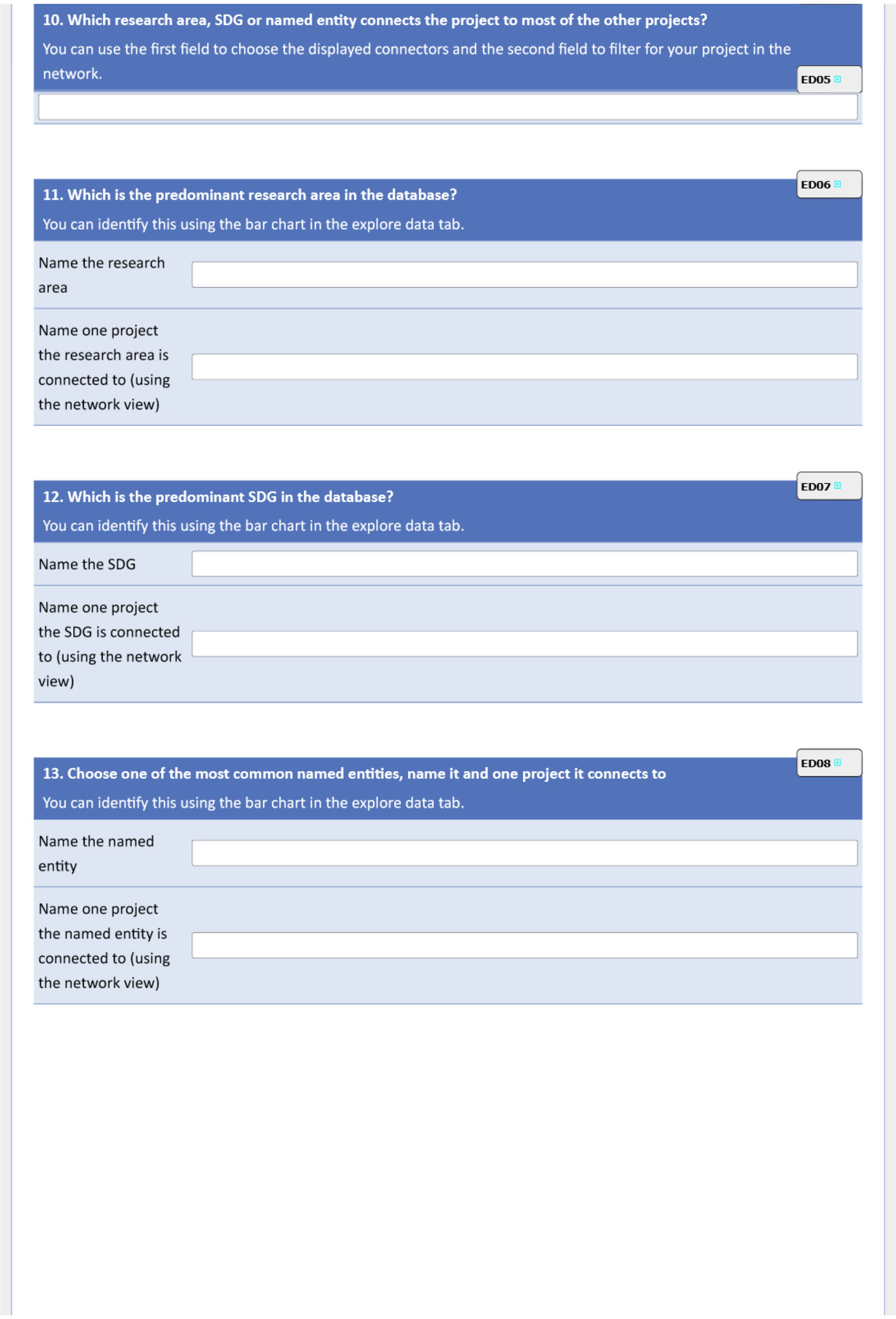
ED03

9. Still using the network view, find two other project it connects to over other connecting nodes (like research areas, SDGs or places)

ED04

01

02



10. Which research area, SDG or named entity connects the project to most of the other projects?

You can use the first field to choose the displayed connectors and the second field to filter for your project in the network.

ED05

11. Which is the predominant research area in the database?

You can identify this using the bar chart in the explore data tab.

ED06

Name the research area

Name one project the research area is connected to (using the network view)

12. Which is the predominant SDG in the database?

You can identify this using the bar chart in the explore data tab.

ED07

Name the SDG

Name one project the SDG is connected to (using the network view)

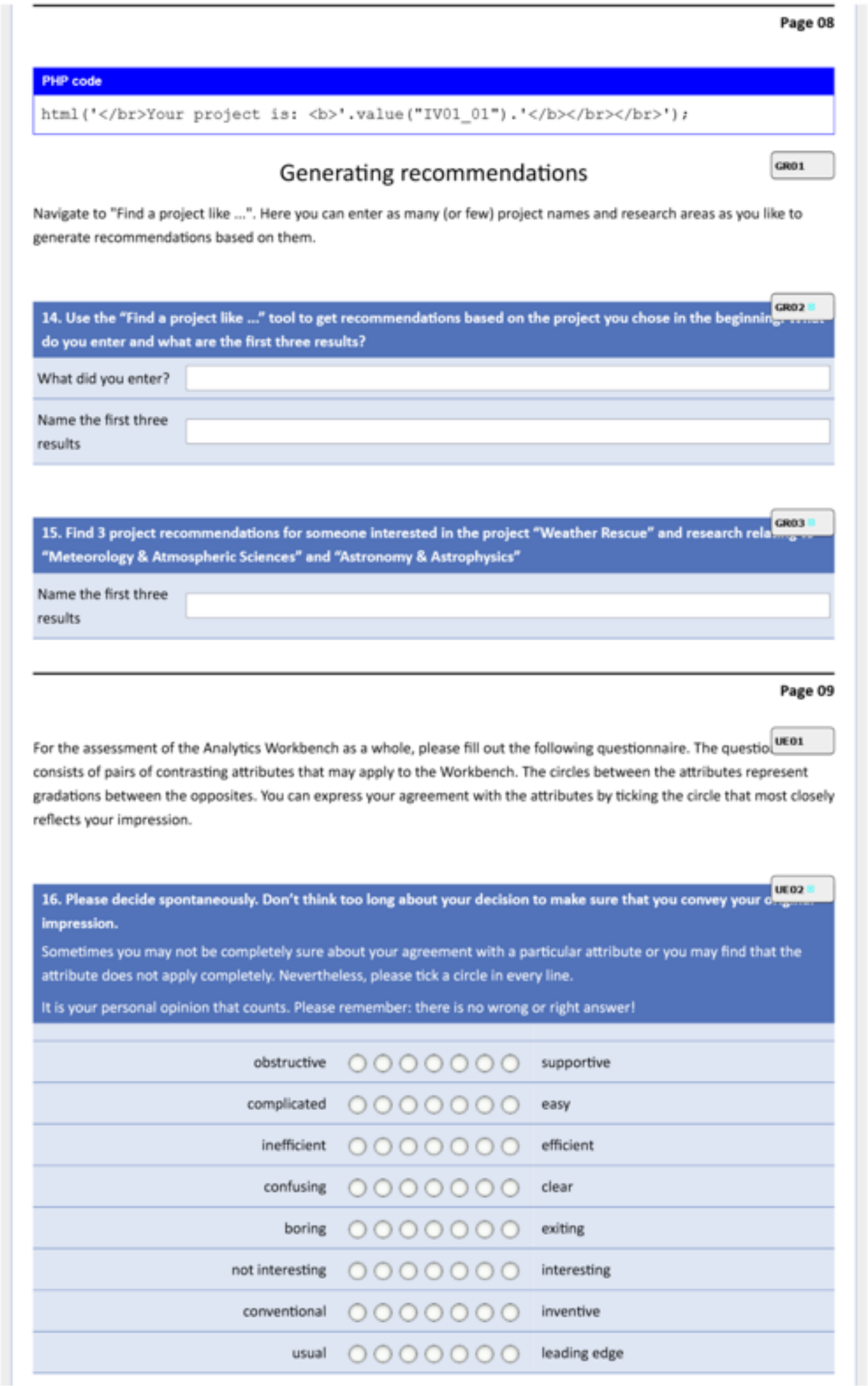
13. Choose one of the most common named entities, name it and one project it connects to

You can identify this using the bar chart in the explore data tab.

ED08

Name the named entity

Name one project the named entity is connected to (using the network view)



Page 08

PHP code

```
html('</br>Your project is: <b>'.value("IV01_01").'</b></br></br>');
```

GR01

Generating recommendations

Navigate to "Find a project like ...". Here you can enter as many (or few) project names and research areas as you like to generate recommendations based on them.

GR02

14. Use the "Find a project like ..." tool to get recommendations based on the project you chose in the beginnin do you enter and what are the first three results?

What did you enter?	
Name the first three results	

GR03

15. Find 3 project recommendations for someone interested in the project "Weather Rescue" and research rela "Meteorology & Atmospheric Sciences" and "Astronomy & Astrophysics"

Name the first three results	

Page 09

UE01

For the assessment of the Analytics Workbench as a whole, please fill out the following questionnaire. The questio consists of pairs of contrasting attributes that may apply to the Workbench. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

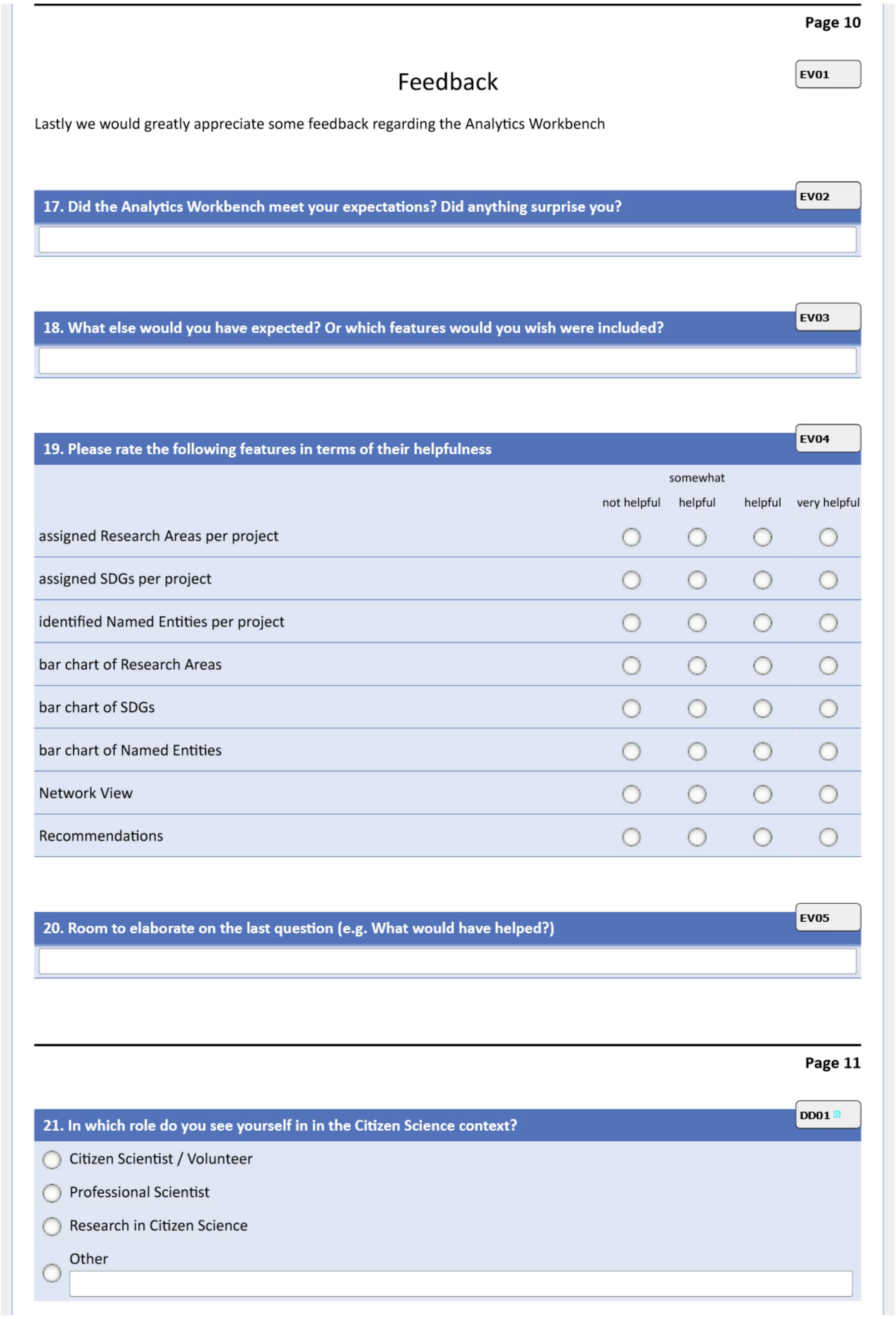
UE02

16. Please decide spontaneously. Don't think too long about your decision to make sure that you convey your o impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

obstructive	○	○	○	○	○	○	○	supportive
complicated	○	○	○	○	○	○	○	easy
inefficient	○	○	○	○	○	○	○	efficient
confusing	○	○	○	○	○	○	○	clear
boring	○	○	○	○	○	○	○	exiting
not interesting	○	○	○	○	○	○	○	interesting
conventional	○	○	○	○	○	○	○	inventive
usual	○	○	○	○	○	○	○	leading edge



Page 10

EV01

Feedback

Lastly we would greatly appreciate some feedback regarding the Analytics Workbench

EV02

17. Did the Analytics Workbench meet your expectations? Did anything surprise you?

EV03

18. What else would you have expected? Or which features would you wish were included?

EV04

19. Please rate the following features in terms of their helpfulness

	not helpful	somewhat helpful	helpful	very helpful
assigned Research Areas per project	○	○	○	○
assigned SDGs per project	○	○	○	○
identified Named Entities per project	○	○	○	○
bar chart of Research Areas	○	○	○	○
bar chart of SDGs	○	○	○	○
bar chart of Named Entities	○	○	○	○
Network View	○	○	○	○
Recommendations	○	○	○	○

EV05

20. Room to elaborate on the last question (e.g. What would have helped?)

Page 11

DD01

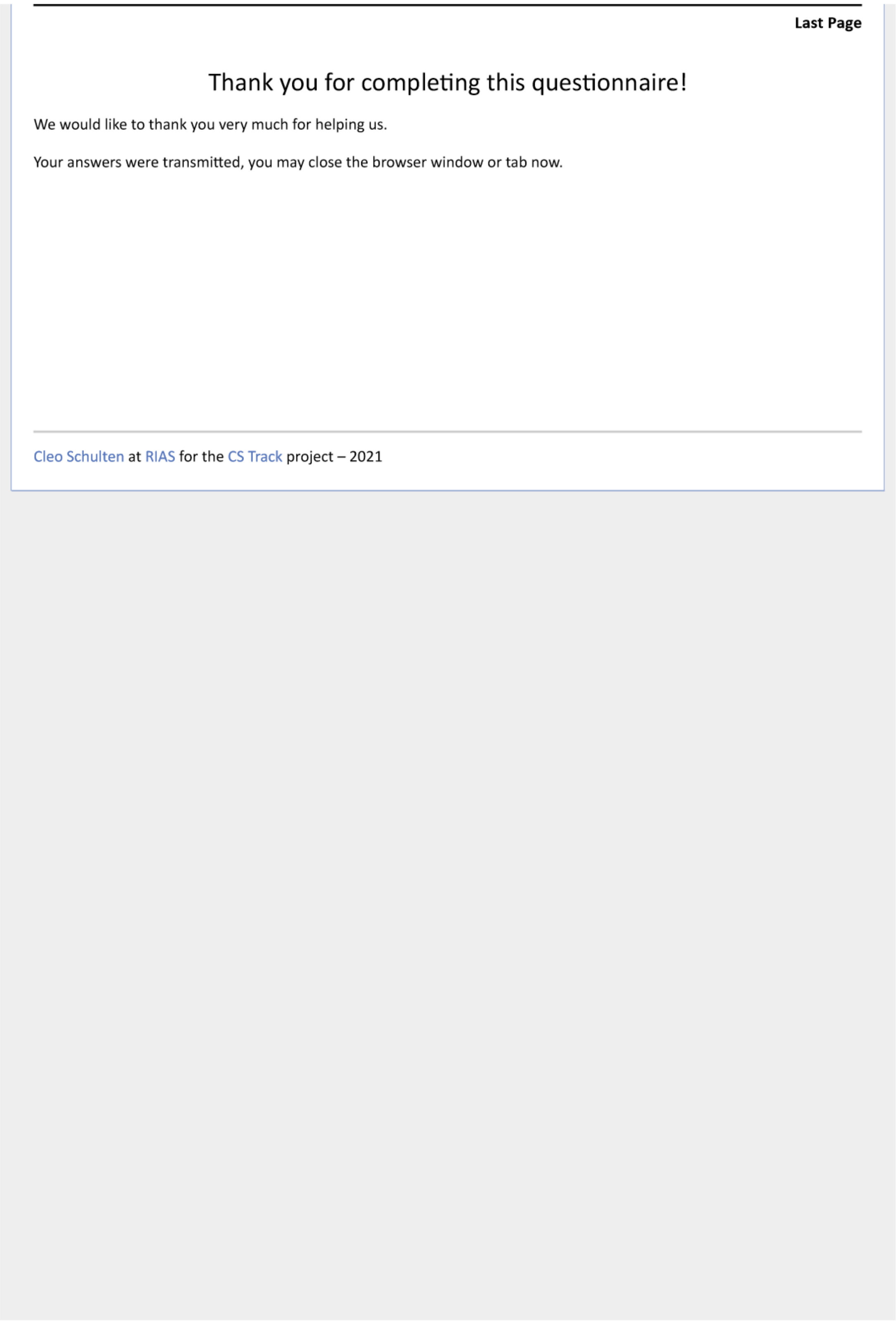
21. In which role do you see yourself in in the Citizen Science context?

○ Citizen Scientist / Volunteer

○ Professional Scientist

○ Research in Citizen Science

○ Other



Last Page

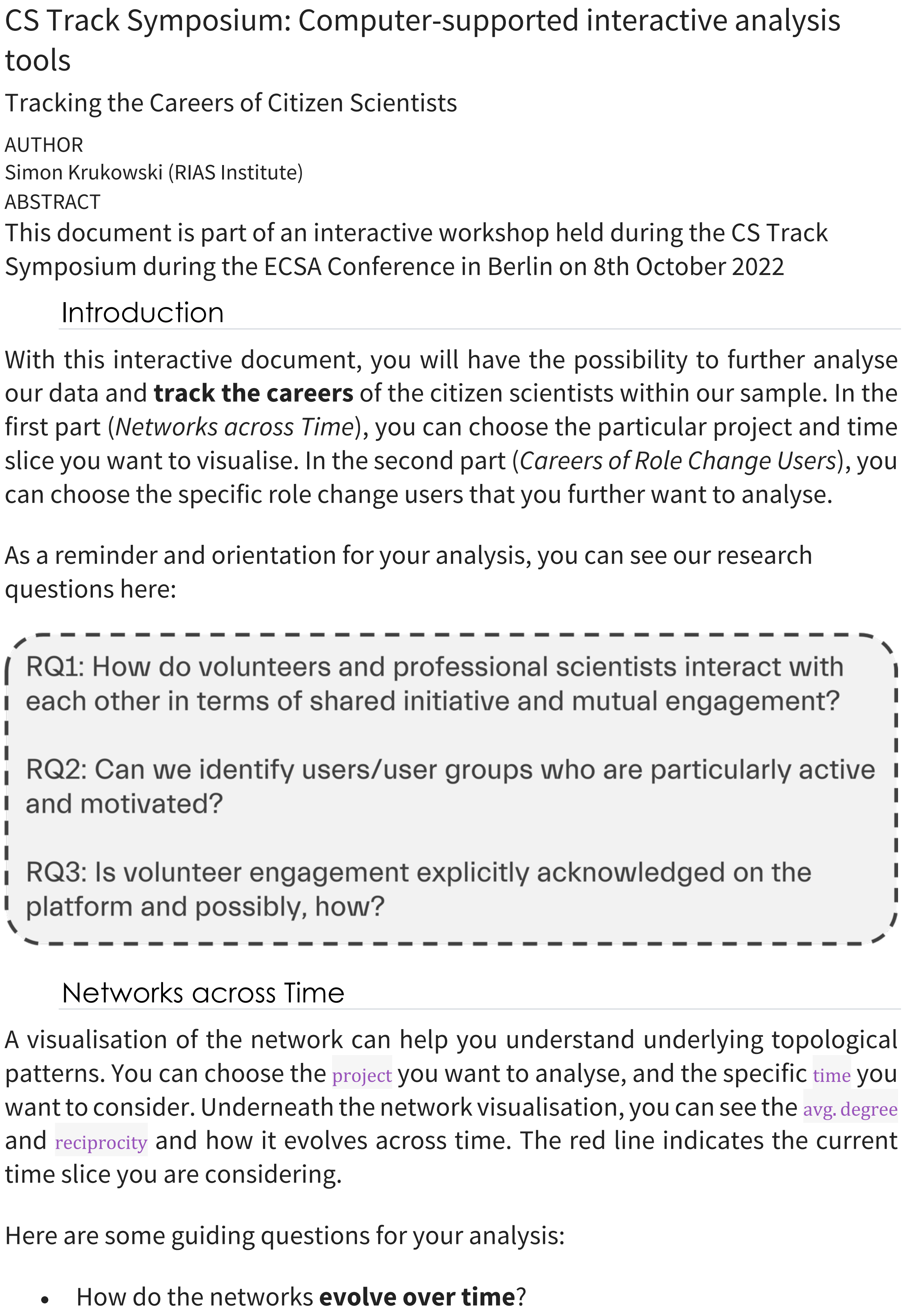
Thank you for completing this questionnaire!

We would like to thank you very much for helping us.

Your answers were transmitted, you may close the browser window or tab now.

Cleo Schulten at RIAS for the CS Track project – 2021

6.2. Snapshot of interactive Quarto document



CS Track Symposium: Computer-supported interactive analysis tools

Tracking the Careers of Citizen Scientists

AUTHOR
Simon Krukowski (RIAS Institute)

ABSTRACT
This document is part of an interactive workshop held during the CS Track Symposium during the ECSA Conference in Berlin on 8th October 2022

Introduction

With this interactive document, you will have the possibility to further analyse our data and **track the careers** of the citizen scientists within our sample. In the first part (*Networks across Time*), you can choose the particular project and time slice you want to visualise. In the second part (*Careers of Role Change Users*), you can choose the specific role change users that you further want to analyse.

As a reminder and orientation for your analysis, you can see our research questions here:

RQ1: How do volunteers and professional scientists interact with each other in terms of shared initiative and mutual engagement?

RQ2: Can we identify users/user groups who are particularly active and motivated?

RQ3: Is volunteer engagement explicitly acknowledged on the platform and possibly, how?

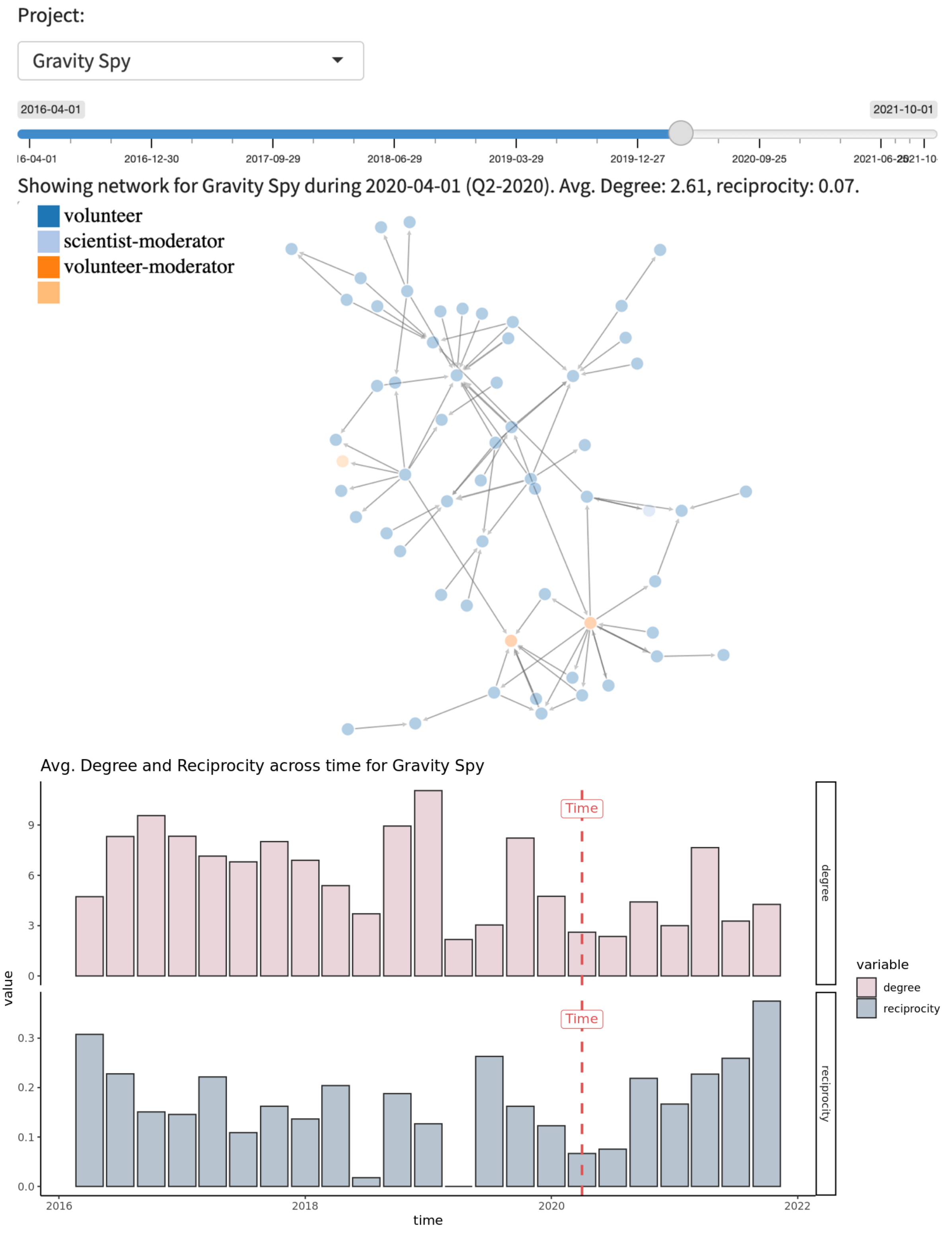
Networks across Time

A visualisation of the network can help you understand underlying topological patterns. You can choose the project you want to analyse, and the specific time you want to consider. Underneath the network visualisation, you can see the avg. degree and reciprocity and how it evolves across time. The red line indicates the current time slice you are considering.

Here are some guiding questions for your analysis:

- How do the networks **evolve over time**?

- Do certain users become **more central**? What about their **user roles**?
- How do avg. degree and reciprocity evolve and how can this be seen in the network?



Project:

Gravity Spy

Showing network for Gravity Spy during 2020-04-01 (Q2-2020). Avg. Degree: 2.61, reciprocity: 0.07.

1.1 Careers of Role Change Users

As explained during the presentation, we identified 14 users who changed their role over the project and got promoted to a higher role (i.e., volunteer-moderator). Here, you have the possibility to explore their individual paths. Below is a reminder regarding the variables:

Variable	Explanation
Degree Rank	Ranks of the degree (i.e., number of connections per user). We calculated ranks to account for fluctuations in general participation. **Low values** (i.e., higher ranks) indicate **more participation**.
In- & Out-Degree	In- & Out-degree reflects the ingoing vs. outgoing connections. In our sample, this reflects the comments made by the user (i.e., replying/answering to a post, out-degree) and the comments the user received (i.e., other users replied/answered to a post, in-degree). High values indicate **high centrality.**
Periodicity	Number of days without any comments (i.e., absence). The mean gives the mean days of absence per quarter, the SD is the standard deviation. Low values indicate **high adherence**.

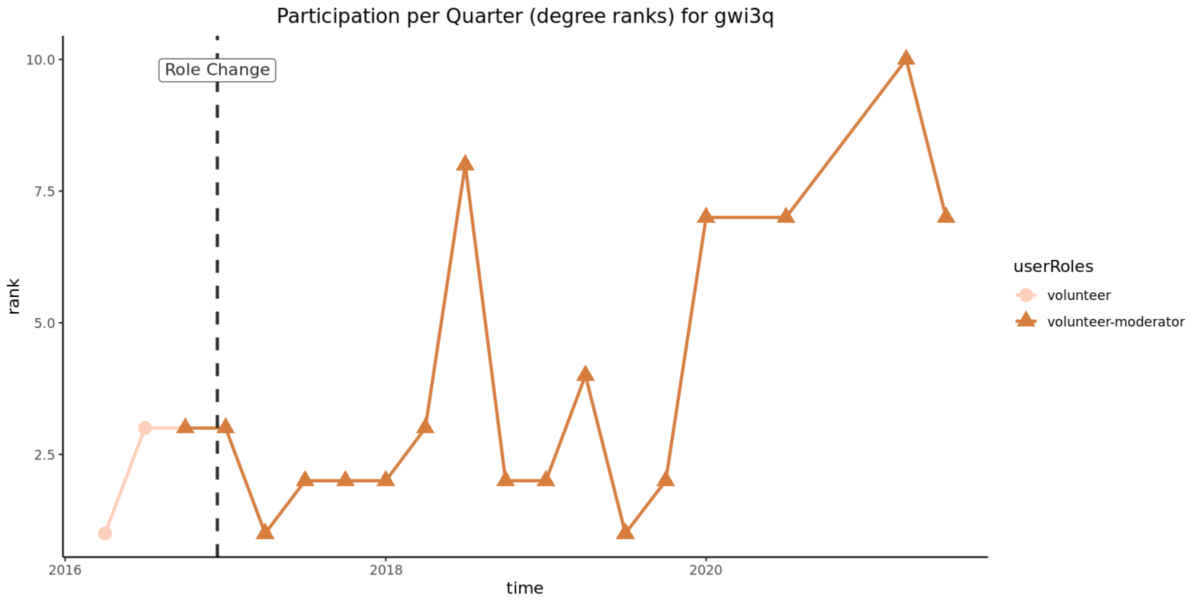
Here are some more guiding questions for your analysis:

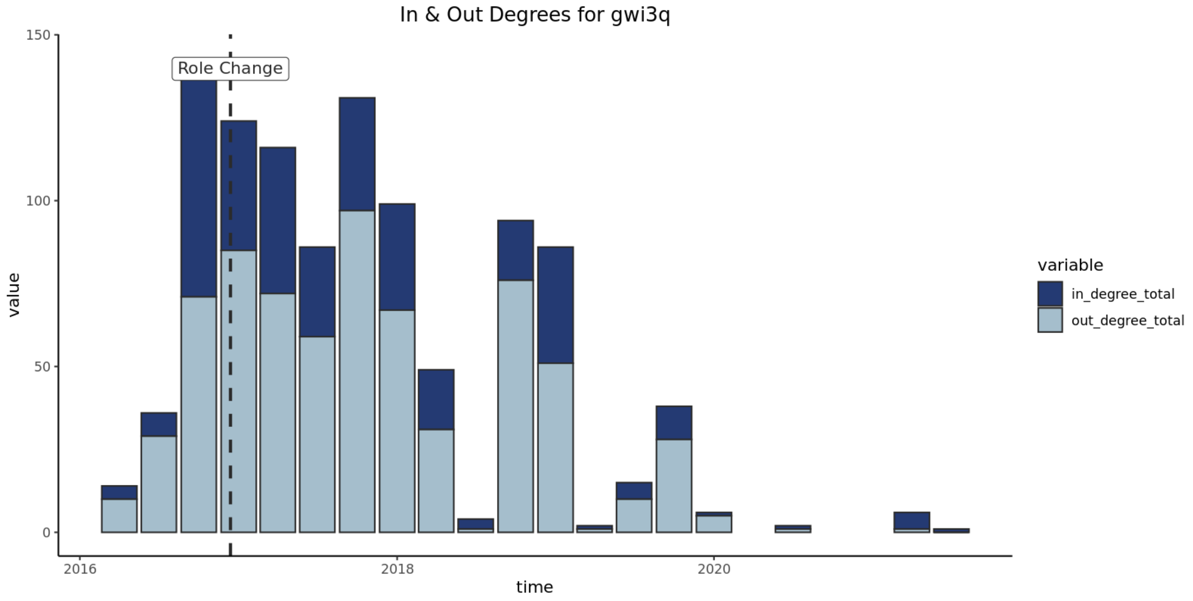
- Can you find users from the above networks in the dropdown (i.e., did they change their role)?
    - What do their careers look like?
- Are there any overarching trends to be observed around the **time of the role change**?

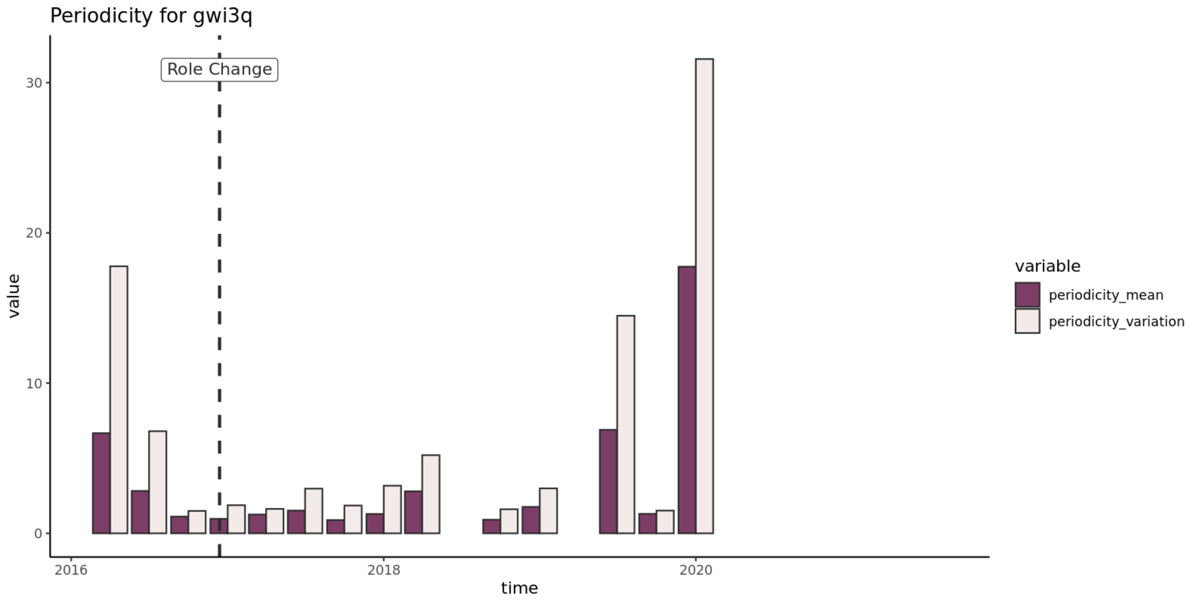
User:

gwi3q

Project	User	Time of role change
Gravity Spy	gwi3q	2016-12-13


Participation per Quarter (degree ranks) for gwi3q
Role Change
rank
10.0
7.5
5.0
2.5
2016
2018
2020
time
userRoles
volunteer
volunteer-moderator


In & Out Degrees for gwi3q
Role Change
value
150
100
50
0
2016
2018
2020
time
variable
in_degree_total
out_degree_total


Periodicity for gwi3q
Role Change
value
30
20
10
0
2016
2018
2020
time
variable
periodicity_mean
periodicity_variation

1.2 Questions?

Visit the project website (https://cstrack.eu) or contact the author (sk@rias-institute.de)