**Proceedings Paper:**
Orphanou, K., Christoforou, E., Otterbacher, J. et al. (2021) Preserving the memory of the first wave of COVID-19 pandemic: Crowdsourcing a collection of image search queries. In: Proceedings of the Third symposium on Biases in Human Computation and Crowdsourcing. Third symposium on Biases in Human Computation and Crowdsourcing (BHCC 2021), 10-12 Nov 2021, Delft, Netherlands (Online). CEUR Workshop Proceedings.

# Preserving the memory of the first wave of COVID-19 pandemic: Crowdsourcing a collection of image search queries

Kalia Orphanou[1], Evgenia Christoforou[2], Jahna Otterbacher[1,2], Monica Lestari Paramita[3] and Frank Hopfgartner[3]

[1]*Open University of Cyprus, Cyprus*
[2]*CYENS - Centre of Excellence, Nicosia, Cyprus*
[3]*Information School, University of Sheffield, United Kingdom*

**Abstract**

The unprecedented events of the COVID-19 pandemic have generated an enormous amount of information and populated the Web with new content relevant to the pandemic and its implications. Visual information such as images has been shown to be crucial in the context of scientific communication. Images are often interpreted as being closer to the truth as compared to other forms of communication, because of their physical representation of an event such as the COVID-19 pandemic. In this work, we ask crowdworkers across four regions of Europe that were severely affected by the first wave of pandemic, to provide us with image search queries related to COVID-19 pandemic. The goal of this study is to understand the similarities/differences of the aspects that are most important to users across different locations regarding the first wave of COVID-19 pandemic. Through a content analysis of their queries, we discovered five common themes of concern to all, although the frequency of use differed across regions.

**Keywords**

Image, Pandemic information, Proprietary search, Crowdsourcing

## 1. Introduction

The COVID-19 pandemic is proving to have a high impact on people's feelings and behavior. People may be suffering from confusion, isolation, and feelings of insecurity [1]. Researchers are citing large-scale problems such as alcohol and drug abuse [2] as well as increased levels of anxiety and sleep disturbances [3]. Given these stresses, it is not surprising that users' COVID-19 related concerns and worries are revealed on what we search for online[1]. Studies have exploited data from Google Trends to identify the most frequent queries during pandemic, examining how search behaviour relates to the epidemic trends using an infodemiology approach [4, 5, 6]. For instance, Canchari et al. [5] used Trends data from searches related to the coronavirus disease during the period of January to May 2020 and identified that "coronavirus" was the

*BHCC 2021*
kalia.orphanou@ouc.ac.cy (K. Orphanou); e.christoforou@cyens.org.cy (E. Christoforou); jahna.otterbacher@ouc.ac.cy (J. Otterbacher); m.paramita@sheffield.ac.uk (M. L. Paramita); f.hopfgartner@sheffield.ac.uk (F. Hopfgartner)


[1]https://www.weforum.org/agenda/2020/05/google-trends-search-online-coronavirus-covid-19/

most frequent search term, followed by “fever,” “sore throat,” and “cough.” In addition, they found that specific queries such as “covid spread,” “face masks,” “stay home,” were related to the increased severity of the pandemic during that period. Using a similar approach, other works [7, 4, 8] utilising both Google Trends and Baidu Index, found out that terms relating to shortness of breath, headache, chest pain and loss of smell correlated with rates of confirmed cases and deaths. Another study investigated aspects influencing the use of specific queries [6]. Whilst they found that the term "coronavirus" was used frequently throughout the pandemic, the use of some queries was found to be influenced by media coverage. E.g., queries such as “ageusia” (loss of taste) and “anosmia” (loss of smell) were only found in the search trends once they have been reported as COVID-19 symptoms by the media.

Our focus is not on conducting an infodemiology study, but rather, we aim to understand the similarities as well as the differences, in terms of the important aspects of the pandemic that users in different geographical locations search for in web image search engines. We focus on *image search*, as a key mechanism for the public to find visual information sources about the COVID-19 pandemic which, without a doubt, shall remain an important event in our collective memory. In our previous work [9], we explored variations in what users “see" concerning the pandemic through Google image search using a two-step approach. The first step was to crowdsource a search task to collect image search queries concerning COVID-19 by inviting participants from four severely affected countries in Europe during the first wave of pandemic (Great Britain (GB), Germany (DE), Spain (ES) and Italy (IT)). In the second step, we used the image search queries collected from the crowdsourcing task to analyze three sources of variation - users’ information needs, their geo-locations, and query language - and study their influences on the similarity of Google image search results. In this work, we focus on the first step, i.e., to provide a detailed description of the crowdsourcing task and analyse the collected data. We address the following research question:

RQ. How similar or dissimilar are the image search queries of people across regions regarding the first wave of the COVID-19 pandemic?

The paper is structured as follows. In Section 2, we describe the methodology used for the collection of the queries. In Section 3, we present the methodology used for the processing of the queries, while in Section 4, we present the analysis of the queries collected. In Section 5, we conclude the paper by discussing the main findings.

## 2. Data Collection

The first step of our methodology is to collect the search queries using a crowdsourcing platform with a large pool of workers established in Europe. The Clickworker[2] platform advertises an attractive population of workers, with 30% being located in Europe. Additionally, it features a function for pre-selecting eligible workers based on the country of residence and gender. To test the platform’s claims and whether we could achieve the desirable distribution of demographic characteristics in the sample, we performed a test run targeting the four countries. Through this process, we also estimated the time required to complete the task, which was ten minutes.

[2] www.clickworker.com/clickworker-crowd/

Following the recommendation of the platform[3] we rewarded workers with 1.60 per completed task according to the above estimation. We then executed four crowdsourcing "campaigns," one for each target country, in which we sought responses from 50 men and 50 women, for a total of 400 participants. Our task was set up as a questionnaire using the template provided by the platform. The task, described in detail below, was presented in English to workers across the four countries, to ensure uniformity. However, workers were encouraged to complete the task in the language that they usually search the Web and we asked them to state explicitly the language of their queries. This decision has not affected to a large degree the language in which crowdworkers reported their queries. We received 26.9% English queries from Germany, 18.9% from Spain and 14.6% from Italy. It is also important to consider that participants who are foreign residents of a given country might have replied in their native language.

Since the goal of this study is to understand the similarities/differences of the image search queries that people across different locations search regarding the first wave of the pandemic, we ran the crowdsourcing tasks during mid-September 2020. The participants were provided with two prompts, and asked to provide three search queries (of up to five words per query) in response to each prompt. They were also told that "you may test your queries in Google Image Search if you wish to check the images retrieved for a given query." Although we did not require this check, it was encouraged to promote quality responses.

The two prompts were as follows:

Prompt 1. The number of photo documentaries that exist depicting the historic pandemic of 1918 is limited. We want to record the Covid-19 pandemic through a photo documentary. Please provide us with three image search queries to search the Web and collect relevant images documenting the current pandemic, and its various dimensions / aspects.

Prompt 2. We want to record the habits that people developed during the Covid-19 lockdown through a photo documentary. Please provide us with three image search queries to search the Web and collect relevant images documenting these habits. You may include examples of both "beneficial" as well as "harmful" habits.

## 3. Data Processing

After collecting the queries, we follow three data processing steps: i) cleaning of the collected queries (*Data Cleaning*), ii) categorization of the queries into thematic categories (*Data Categorization*) and iii) aggregation of the queries into similarity groups (i.e., representative queries) (*Data Aggregation*), as detailed in the following subsections.

### 3.1. Data Cleaning

As in similar crowdsourcing tasks, we face the issue of the quality [10] of the collected results. In this respect, we conducted one mitigation method and two auditing methods to help us ensure the quality of the queries, i.e., that the queries used throughout the work are of quality in terms of matching the task requirements and the conceptual requirements of this work (i.e., are as realistic as possible).

[3] https://www.clickworker.com/survey-participants-for-online-surveys/fee-recommendations

As a mitigation method to remove spam, we removed any participant responses that had an overall low quality (i.e., in all six required queries). In the context of our task, low quality responses are ones that provided a link instead of a query, or a reply that is completely out of topic (i.e., '1918 flu pandemic'). In total, we discarded 50 responses out of the original 400. This initial data cleaning still allowed us to maintain a fairly balanced dataset in terms of country and gender representation (see Table 1 for details), which was our initial objective (50 men and 50 women per country).

**Table 1**

Gender of the Crowdworkers.

| Country \| Gender | Male | Female | Total |
|---|---|---|---|
| GB | 48 | 48 | 96 |
| IT | 46 | 46 | 92 |
| ES | 35 | 45 | 80 |
| DE | 39 | 43 | 82 |
| Total | 168 | 182 | 350 |

Furthermore, to audit the dedication of the participants to the task [10], and verify that our payment was fair, we looked at the time it took each participant to respond to the task (i.e., task duration). For participants in our sample, the median task duration was 9.7 minutes and the most frequent task duration was 3.3 minutes. On the other hand, the duration of the discarded responses had a median task duration of 6.8 minutes and the most frequent task duration was 1 minute. Given the median of the discarded responses we have an additional indication that we fairly discarded those responses, as being low quality.

Finally, as a last measure to audit the appropriateness of collected results compared to the objectives of this work, we asked participants to indicate how frequently they used the image search function. 40% of participants self-reported that they used image search daily while 44% reported that they used it 1-3 times per week. Additionally, 25% of participants reported that image search was their principal source of information and for 73% of the participants it was their secondary source. These results show that participants are a representative sample of image search users, who frequently use image search as an essential source of information. For a detailed report on how participants per country use the image search function, see Figure 1 and Figure 2.

The queries from both tasks were merged, and then cleaned and tokenized, creating a gender-balanced dataset for each of the four countries. Steps in the query cleaning were as follows: i) replacement of tab, newline and multiple space characters with a single space; ii) all text were converted to lower-case; iii) all the expressions referring to 'covid' e.g. 'corona' or 'coronavirus', were replaced with the word 'covid'; iv) any URLs were removed. We identified all the unique queries collected from the participants in each country and computed the number of appearances for each unique query without considering any duplicates of the same user, i.e. frequency.

## 3.2. Categorisation of the queries

Next, we aimed to categorise the queries in terms of the users' information needs; thus, we performed a manual content analysis using an inductive approach. Devising appropriate

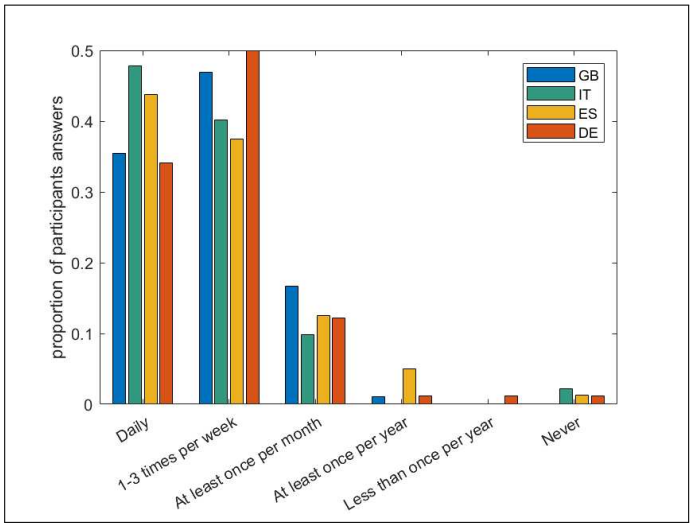


**Figure 1:** Proportion of participants' answers to the following question: "How frequently do you use the image search function to search on the Web?"

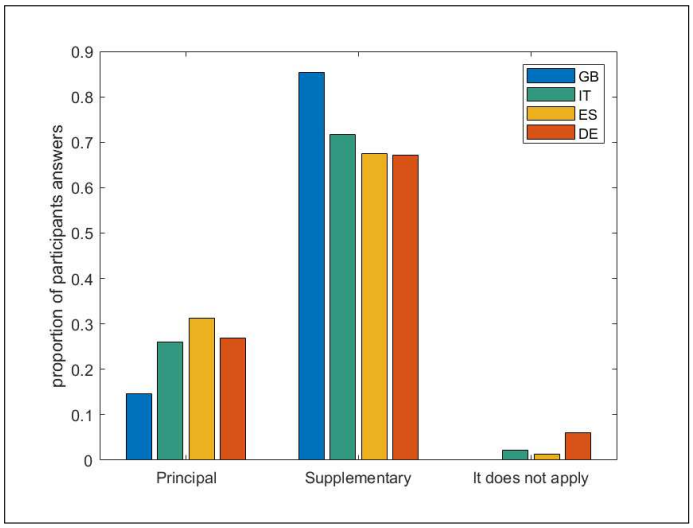


**Figure 2:** Proportion of participants' answers to the following question: "The image search as a source of information for me."

categories for the queries is part of our methodology in order to answer our RQ. In other words, we created a base of comparison among the various queries received from crowdworkers across regions.

Initially, three researchers examined the GB queries, discussing the topics expressed in them, until a consensus on six categories was reached. Next, two researchers analysed the remaining queries from the other locations (DE, ES and IT), involving a third researcher to resolve any disagreements. We were careful to consider whether or not additional categories might be needed, given that the six categories were defined through the analysis of the GB data; however,

it was found that the taxonomy was applicable across the data from all regions. Each query was mapped onto one and only one category. The categories defined through the content analysis, are as follows:

- **Stay at home:** Queries affirming or asking about habits or actions while in a stay-at-home restriction or lockdown. These queries describe the habits developed due to the stay at home restriction. Also, it includes queries stating the impact of the pandemic to a person's mental state and well-being. *Examples:* "`covid zoom call`", "`covid food delivery`".
- **Personal Protection**: Queries asking or describing a personal protection instruction or measure during the pandemic. If the concept of the query can be interpreted as personal protection measure it is included here; thus, queries about equipment or accessories needed are included as well. This category also hosts queries asking general questions about the "do's and don'ts" during the pandemic. *Examples:* "`face mask`", "`hand washing`".
- **Healthcare**: This category hosts queries relevant to the healthcare system, the way it was impacted, and the means/methods for identifying COVID-19. *Examples:* "`covid vaccine`", "`covid test centre`".
- **Pandemic General Information**: General queries regarding the pandemic, e.g., how much it has spread in the world and queries asking for statistical facts. Includes queries asking about covid in certain geographical areas. *Examples:* "`covid outbreak`", "`covid in Italy`".
- **Society/Community Impact**: Queries asking or describing the impact that the COVID-19 pandemic & measures had in the society and the different communities (i.e., at a collective level). This category includes general queries relevant to social phenomena in time and space that were not present before. *Examples:* "`covid empty streets`","`covid NHS clap`".
- **Miscellaneous**: This category is used for any queries that do not fall into any of the above and/or of which the meaning cannot be clearly interpreted. *Examples:* "`1918 flu pandemic`", "`5 edtech startup`". For the purpose of this study, we do not analyze the queries from this category.

### 3.3. Aggregation of the queries

After categorising the queries, we identified that some queries were very similar but contained differences in the word order, synonyms or including the word "`covid`". Initially, we considered the queries collected from participants for each of the four countries independently. To produce a more robust analysis, for each category, we merged similar queries together if they contain: i) the same words but in different order; ii) synonyms; iii) the same words with "`covid`", "`image`", or "`photo`". Queries that are sub-sets of each other, e.g., "`covid hospital`" and "`covid nightingale hospital`" were considered to be separate. From each merged group, we selected the most frequent query as the representative query, the frequency of which represents the total frequency of the queries in the group (see Table 2).

As a further step, we merged the queries collected from all the four countries using the same aggregation procedure as before. In addition, we merged similar queries in different languages

**Table 2**

Example of a GB merged group and its representative query

| Queries | Freq | Representative query | Freq |
|---|---|---|---|
| covid symptoms | 11 | covid symptoms | 17 |
| symptoms of covid | 2 | | |
| covid virus symptoms | 2 | | |
| covid symptoms factsheet | 1 | | |
| symptoms | 1 | | |

if they contained: i) the exact translation of the query, ii) synonyms, iii) the translation of the query in another language with the word: "covid", "image", or "photo". As in the first step of aggregation, from each merged group, we selected one representative query, the frequency of which represents the total frequency of the queries in the group (see Table 3).

**Table 3**

Example of merged group after merging all the queries from all the countries and its representative query.

| Queries | Freq | Country | Translation | Representative | Freq |
|---|---|---|---|---|---|
| covid symptoms | 11 | GB | | covid symptoms | 30 |
| symptoms of covid | 2 | GB | | | |
| covid virus symptoms | 2 | GB | | | |
| covid symptoms factsheet | 1 | GB | | | |
| symptoms | 1 | GB | | | |
| covid krankheitszeichen | 1 | DE | signs of illness | | |
| covid symptome | 4 | DE | covid symptoms | | |
| covid anzeichen | 1 | DE | covid signs | | |
| covid symptoms | 2 | ES | | | |
| covid symptoms photos | 2 | ES | | | |
| covid sintomi contagiati | 1 | IT | covid symptoms infected people | | |
| covid sintomi malattia | 1 | IT | covid disease symptoms | | |
| sintomi covid virus | 1 | IT | covid virus symptoms | | |

## 4. Analysis of the Collected Queries

Here we address RQ – *How similar or dissimilar are the image search queries of people across regions regarding the first wave of the COVID-19 pandemic?* The queries collected from the participants were categorized into one of five thematic categories and then aggregated into similarity groups (i.e., representative queries), as detailed in Section 3.2. We compare the distributions of the thematic categories, as they are used across countries, considering the number of representative queries.

The results presented in Figure 3 provide an overall comparison of the *variety of queries*

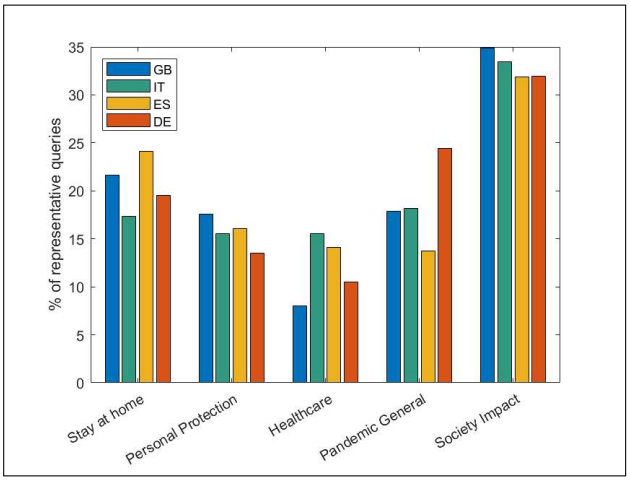


**Figure 3:** Percentage of the representative queries per category and country.

collected from each country in each category by computing the percentage of representative queries for each category over the total number of representative queries per country (Table 4). Before going into the analysis of the results, we performed a chi-square test of independence to examine the relation between the categories and the country of residence for the set of representative queries. The relation between these variables was significant, $X^2(15, 1223) = 24.0308$, and $p = 0.02$. For this reason, we have a closer look to this relationship.

**Table 4**

Number of representative queries by category and country. In parenthesis is the total number of queries in each category. 'Merged' represents the total number of queries on each category after merging all the queries from all the countries.

| | **Stay at Home** | **Personal Protection** | **Healthcare** | **Pandemic General** | **Society Impact** |
|---|---|---|---|---|---|
| ES | 72 (87) | 48 (80) | 42 (60) | 41 (74) | 95 (126) |
| DE | 52 (74) | 36 (58) | 28 (34) | 65 (105) | 85 (117) |
| IT | 58 (84) | 52 (80) | 52 (77) | 61 (91) | 112 (153) |
| GB | 70 (108) | 57 (83) | 26 (43) | 58 (89) | 113 (153) |
| Merged | 183 (353) | 108 (301) | 121 (214) | 144 (359) | 288 (549) |

In the *Stay at Home* category, the crowdworkers in Spain provided the highest percentage of representative queries which means that they provided diverse topics of queries in this category, in contrast to the Italian crowdworkers who reported the lowest variety. The German participants reported 4.6% less representative queries compared to the Spanish sample, where as the Great Britain sample had the second highest percentage of representative queries. Considering the *Personal Protection* theme, we notice that the German sample reported the lowest percentage of representative queries, thus the highest similarity of queries while the GB sample had the opposite behaviour, having the most diverse queries.

In contrast to the rest of the categories, in the *Healthcare* theme, the GB crowdworkers

reported less representative queries than in other categories. On the other hand, Italian crowdworkers submitted the highest percentage of representative queries in this category which can be justified by the fact that the Italian population had a larger time of exposure than in other countries to the pandemic during the time of the study, and this is reflected in at least in the *Healthcare* category that includes queries relevant to hospitals, symptoms, etc.

Regarding the *Pandemic General Information* category, German crowdworkers have a much larger percentage of representative queries (i.e., lowest similarity), at least 6.2% more compared to the rest of the countries. This is the largest difference of representative queries among countries over all the categories. It appears that German crowdworkers are "preoccupied" with various general topics that relate to the pandemic and they have a more diverse way of expressing those queries compared to workers from other countries. Moreover, the Italian and GB sample of queries have a very similar percentage of representative queries ranking second and third respectively, compared to the other countries.

As per the *Society Impact* category, for every country, the percentage of representative queries is the highest among all the categories. This indicates that every sample of queries from each country has a higher focus on describing social phenomenon emerging from the pandemic or the subsequent effects of the measures to restrict the spread of the virus. Queries reported by the Spanish and German workers in this category have almost the same percentage of representative queries and the lowest among all countries regarding Society Impact. The sample of queries received from GB on this category was the most diverse expanding to different topics (highest percentage of representative queries).

## 5. Conclusion

In this work, we collect image search queries through crowdsourcing to find out what kind of visual information do users search for on the web during the first wave of COVID-19 pandemic. This allows us to have access to a diverse set of people with "web literacy" and collect a wide range of queries of diverse topics, that we wouldn't have the possibility to identify only by looking at popular COVID-19 queries. Using content analysis, we created a taxonomy of five common themes for categorising all user queries from four locations. We found clear evidence that, although participants' queries related to five common themes, the frequencies with which people mentioned particular themes, and the extent to which there was a rich variety of queries within a theme, varied by their country of residence. This is, of course, a reflection of experiences, given that the pandemic has played out very differently across regions, even within Europe. As concrete evidence, Italian participants, whose country has been the most severely affected during the first wave of COVID-19, had the largest number of Healthcare queries, as well as the richest vocabulary of such queries. However, one commonality across regions, was that participants' queries often focused on Social Impact, or new social phenomena experienced by all of us and less on healthcare. As a future direction, we will examine the behaviour of crowdworkers in formulating image search queries based on their cultural background. To this end, we plan to enlarge our dataset, and expand on the idea of how the demographic

differences (i.e. gender, age, language, nationality) among workers are associated or can explain the provided queries and result differences.

## References


[1] B. Pfefferbaum, C. S. North, Mental health and the covid-19 pandemic, New England Journal of Medicine (2020).

[2] J. M. Clay, M. O. Parker, Alcohol use and misuse during the covid-19 pandemic: a potential public health crisis?, The Lancet Public Health 5 (2020) e259.

[3] L. Sher, Covid-19, anxiety, sleep disturbances and suicide, Sleep Medicine (2020).

[4] A. Walker, C. Hopkins, P. Surda, Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak, International Forum of Allergy & Rhinology 10 (2020). doi:https://doi.org/10.1002/alr.22580.

[5] C. R. A. Canchari, S. G. Chávez-Bustamante, B. S. Caira-Chuquineyra, Exploratory analysis of internet search trends during the COVID-19 outbreak (2020).

[6] B. Sousa-Pinto, A. Anto, W. Czarlewski, J. M. Anto, J. A. Fonseca, J. Bousquet, Assessment of the Impact of Media Coverage on COVID-19–Related Google Trends Data: Infodemiology Study, Journal of Medical Internet Research 22 (2020) e19611.

[7] T. S. Higgins, A. W. Wu, D. Sharma, E. A. Illing, K. Rubel, J. Y. Ting, S. F. Alliance, Correlations of Online Search Engine Trends With Coronavirus Disease (COVID-19) Incidence: Infodemiology Study, JMIR Public Health and Surveillance 6 (2020). doi:10.2196/19702.

[8] H. C. Cousins, C. C. Cousins, A. Harris, L. R. Pasquale, Regional infoveillance of covid-19 case rates: Analysis of search-engine query patterns, Journal of Medical Internet Research 22 (2020) e19483. doi:10.2196/19483.

[9] M. L. Paramita, K. Orphanou, E. Christoforou, J. Otterbacher, F. Hopfgartner, Do you see what i see? images of the covid-19 pandemic through the lens of google, Information Processing Management 58 (2021) 102654. doi:https://doi.org/10.1016/j.ipm.2021.102654.

[10] U. Gadiraju, R. Kawase, S. Dietze, G. Demartini, Understanding malicious behavior in crowdsourcing platforms: The case of online surveys, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 1631–1640.


## Acknowledgments


This project is partially funded by the Cyprus Research and Innovation Foundation under grant EXCELLENCE/0918/0086 (DESCANT) and by the European Union's Horizon 2020 research and innovation programme under grant agreements No. 810105 (CyCAT) and 739578 (RISE).