CITIZEN SCIENCE:
THEORY AND PRACTICE

Rohden, F, et al. 2019. Tagging, Pinging and Linking – User Roles in Virtual Citizen Science Forums. *Citizen Science: Theory and Practice*, 4(1): 19, pp.1–13. DOI: https://doi.org/10.5334/cstp.181

RESEARCH PAPER

Tagging, Pinging and Linking – User Roles in Virtual Citizen Science Forums

Frauke Rohden, Christopher Kullenberg, Niclas Hagen and Dick Kasperowski

This article investigates user roles in virtual citizen science projects through a case study of the Talk-forum of Shakespeare's World, a humanities project on the Zooniverse platform. To address collective knowledge production, we study the use of hashtags, pinging, and linking as a way of differentiating how researchers, moderators, and different user groups use the forum according to their roles. We show how both volunteers and researchers have a much deeper interest in the texts that they transcribe and actively seek contextual information, shape new lines of inquiry, and discover new phenomena. We conclude that the use of online forums in citizen science can play a crucial role for extending the knowledge production from academic research to a wider public interest, and also provide new knowledges beyond the assigned task of transcribing documents.

Keywords: citizen science; citizen humanities; discussion forum; Virtual Citizen Science

Introduction

The rapid adoption of information and communication technologies has created several new opportunities for researchers and volunteers to create knowledge collaboratively. Ranging from web-based solutions to mobile applications, several ways of discussing, disseminating, and classifying information have been introduced, opening up new interfaces that barely existed only a decade ago. Contemporary online environments, from the social media giants to forums for niche interests, offer several technological functionalities that increase the prospects for communication among communities and topics, for example by tagging information, notifying users through "pings," and using hyperlinks to contextualise and connect instances of information. From the perspective of citizen science, understanding how researchers and volunteer contributors interact online and create knowledge typology together become key issues. Not only is this understanding valid for discovering new configurations between different forms of expertise, it also is important for improving citizen science as a research method that can be used over a wide range of scientific fields of inquiry. In this article we will discuss some ways in which researchers and volunteers collaborate and communicate with each other. We will investigate linking, pinging, and tagging behaviour among different user groups. This will contribute to a better understanding of the co-production of knowledge between scientists and non-scientists in VCS.

A few notable studies about online citizen science have emerged, and the interest in what forms of knowledge that citizen scientists generate on discussion forums has been sparked by a few scholars in recent years. In a study by Tinati et al. (2015), the authors traced the development of the Zooniverse platform and derived design recommendations for future projects that included the creation of the Zooniverse "Talk" discussion forums, which tie in with the platform's tasks and allow object-specific discussions. They note that experienced users can take over more advanced roles, for example, as moderators filtering questions and notifying the science team if the community cannot answer a question by itself. In another study, Liberatore et al. (2018) analysed what they refer to as "communities of practice" in a Facebook group administered by researchers. The authors found that the users "used the group to share excitement, ideas, and knowledge about New Zealand garden birds …" (Liberatore et al. 2018: 11).

Moreover, some studies have performed data-driven analyses of VCS discussion forums (e.g., Luczak-Roesch et al. 2014; Ponciano and Brasileiro 2014), finding that VCS often has a small number of users contributing a large percentage of the activity in the project. Hedges and Dunn (2017) call such users "super-contributors" in contrast to the larger number of users who make fewer contributions in total, often only one or two contributions each. Reed et al. (2013) investigated the number of minutes spent on classifying data in the Galaxy Zoo project and used this distribution to form a stratified sample of users, which could then become the target for their survey on motivation. In a similar vein, Ponciano and Brasiliero

University of Gothenburg, SE
Corresponding author: Christopher Kullenberg (christopher.kullenberg@gu.se)

(2014) used four different metrics of user activity to distinguish "engagement profiles" among the users in the projects Galaxy Zoo and The Milky Way Project. However, data-driven approaches, in which user contributions are divided into segments with a quantitative approach, have both benefits and drawbacks. While they identify devoted users who spend a large amount of time in the project and provide a good measure of the contributions made by the so called "long tail" (i.e., a large number of users that make only a few contributions each), they are insensitive to the particular roles that have functional differences in a discussion environment, for example as experienced or expert users. In our case, such roles are moderators and researchers. The roles of moderators in a forum discussion will not be detected by quantitative measures alone, because such roles may show similar patterns as both super contributors and, on the other side of the spectrum, casual users. Yet, the moderator's functional roles can be crucial for the dynamics and knowledge production in a VCS forum. For example, a researcher will bring in specific expertise, often valued as authoritative, while moderators have a duty to keep discussions on topic and resolve emerging issues and conflicts. These qualities are not readily detected through dividing user contributions in segments with a quantitative approach.

Furthermore, there are non-linear aspects of forum environments that cannot be quantified in a simplistic manner. By non-linear, we mean that affordances and functionalities in VCS forums have bearing on the technical structure of the forum offering users multiple ways of interacting with the forum. Hashtags (#), introduced on multinational platforms such as Facebook, Twitter, and Instagram have entered the realm of citizen science. Studies have shown the importance of hashtags as markers of content, symbols of community, and "influencers" (Zhang, Zheng, and Pang 2018) and, in terms of learning, reflecting users' needs in relation to hashtags (Veletsianos 2017). Patterns created by hashtags have been found to be both stable and replicated as well as allowing for minority opinions (Golder and Huberman 2006), as hashtags often can be used without constraints. However, they also develop into tag clouds marginalizing groups and individuals (Sinclair and Cardew-Hall 2008), because the visualisation algorithms give preference to more commonly used hashtags. On Twitter, research has shown that scientific authority is likely to lead to virtual authority, and the "hashtagging habits" of researchers participating in scientific conferences have been found to mainly be directed at peer researchers (Letierce et al. 2010). Studies of museum curatorial contexts have shown that terminology of tagging by contributors differs from that of professional curators (Trant 2009, Trant, Bearman, and Chun 2007), implying different user behaviours. With regards to citizen science, Hedges and Dunn (2017) have pointed out that collaborative hashtagging serves as an important aspect of organising co-productive activities.

For example, on the Zooniverse project Shakespeare's World, there is a clear distinction between tasks for the volunteers, who transcribe text from scanned images of individual pages, and the researchers, who analyse the refined and compiled data. However, the discussion forum, called "Talk," allows for a second, more open and collaborative form of knowledge generation where volunteers can discuss interesting findings amongst each other and with the project's researchers. This form of knowledge generation has led to several discoveries made by volunteers on the Zooniverse platform (e.g., Tinati et al. 2015: 4072). Kasperowski and Hillman (2018) have found that the discussion forum of the oldest Zooniverse project, Galaxy Zoo, contains not only discoveries made by volunteers but also reveals tensions between project staff and volunteers as the volunteers develop their own interests. Thus, it becomes a crucial issue to further investigate the impact of VCS forums with regards to co-creation of knowledge, also between scientists and non-scientists.

Thus, an interesting area of study emerges as online forums can be investigated to know more about how scientists and non-scientists interact to co-create knowledge. The few existing studies in this field point to a small number of highly engaged users in the discussion forums (Kasperowski and Hillman 2018; Luczak-Roesch et al. 2014; Ponciano and Brasileiro 2014; Tinati et al. 2015), but we know very little about the social dynamics between researchers, moderators, and volunteers, and how they create knowledge *together*. With the rapid digitization of citizen science, and the growth in popularity for online platforms, it is necessary to know more about the interactions on these platforms, both for understanding the relationships between researchers and the public and to inform the future design of citizen science projects.

Purpose and Research Questions

The purpose of this study is to analyse how a VCS discussion forum is used, with particular focus on user roles in relation to co-creation of knowledge. We intend to fulfill this purpose by investigating user behaviour in the Shakespeare's World discussion forum in relation to the conventions and technical features offered by the platform. This includes the use of hashtags (#), the use of pinging (notifying users about discussions, so-called @-messages), and the use of external resources, which in turn offer new ways of systematising knowledge and discussion on VCS forums. We will structure our analysis according to the following research questions:

1. How are user contributions distributed among different roles?
2. In what way are external knowledge resources–for example, links (hypertext)–used to bring in external information?
3. How are hashtags used and distributed over time?
4. How are pings (@-messages) used to interact and generate knowledge?

Methods and Materials

Shakespeare's World is a citizen science project hosted on the Zooniverse platform. In this project, participants are invited to transcribe historical documents written by Shakespeare's contemporaries, allowing both researchers and volunteers to learn more about the historical context

of Shakespeare's work. Additionally, the project aims to record new words and word variants for the Oxford English Dictionary (OED).

We have chosen a research design that combines qualitative and quantitative analytical approaches. This design makes use of both forum posts (qualitative) and accompanying meta-data (quantitative). Because these forums are distributed and the users can be scattered around the world, conventional ethnography is not feasible. However, online platforms offer digital "traces" in the form of meta-data that can be used to reconstruct the interactions in detail. Geiger and Ribes (2011) call this approach "trace ethnography." The benefits of this method, compared to interviews or questionnaires, include: 1) All users who have been active on the forum are included in the sample, 2) it is possible to quantify differences between users and user groups, 3) meta-data are more accurate than the individuals' recollection of events, and 4) interviewer bias is avoided.

Concerning the validity of the study, the main challenge was to define various user roles in the discussion forums in relation to actual differences in online behavior. On the one hand, there are pre-defined roles marked in the forums for "moderators" and "researchers." However, the rest of the users are undifferentiated in the platform. As discussed above, research has shown there are good reasons for making further distinctions based on the degree of activity. We chose to draw a very simple quantitative distinction between three groups of users.

Concerning the reliability of the study, the research design has inherent strengths and weaknesses. Forum data are recorded by a computer, thus avoiding mistakes of omission or inaccuracy. However, data are created by human users with a degree of variance in spelling and linguistic conventions. The challenge here is to classify natural language with computational methods. This is most crucial when search expressions are used to classify data. On the one hand, possible spelling and grammar variants have to be taken into account (e.g., #catholic and #catholics or difficult to spell usernames), while on the other hand, the search functions of different software produce different results depending on the internal programming of their regular expression engines, which in turn are used to search for character strings. We accounted for such software bias by running the same search queries through two different softwares (Microsoft Excel and Python).

The data collected consists of 11,450 posts from the project's online discussion forum "Talk," covering the activity within the first two years of the project (2015-11-09 to 2017-11-21). The data were exported using the built-in administrator function of the Talk software as a .json file, which was subsequently converted into various formats for analysis, such as spreadsheet files and textual corpora for rapid and specialised searches. The data were analysed using filtering functions and charts in Microsoft Excel®, the Python programming language including a range of software libraries. For social network analysis (SNA), we used the data visualisation software Gephi (Bastian et al. 2009) (for code and classificatory protocols, see the following Github repository: https://github.com/christopherkullenberg/talk-analyzer).

Scheme of analysis and methodological approach

The Zooniverse Talk forum has a number of technological affordances that shape the departure point of our investigation. These are summarized in **Table 1**.

These functionalities form the software basis for the Talk forum, and we have used them as ways of instantiating our line of inquiry. However, we will study how these functions are used in practice rather than taking for granted that they are used as intended or programmed. We will proceed along similar lines as Hedges and Dunn note, that "the rise of social media, especially multinational platforms such as Facebook, Twitter, and Instagram, has introduced the hashtag into the crowd's daily consciousness and has had a significant effect on the dynamics of tagging" (2017: 34). In other words, by extending this analysis to include pings, threads, and linking practices, we have structured the quantitative parts of our analysis on the extraction of empirical indicators as outlined in **Table 1** by quantifying the length of threads, the frequency of hashtags, the frequency of @-messages, and the frequency of links to webpages. Moreover, we have created two other data structures–one that is temporal, in which we study the development of a hashtag over time, and one that is of a network character–to connect on a user-to-user basis the pinging practice in various hashtags. These types of data are necessary to properly address our research questions (RQ). In particular, RQ 3 and 4 require more than just frequency measures; they must also include timestamp data and, in the latter case, the creation of network structure data.

Ethical considerations

The Talk forum user agreement (https://www.zooniverse.org/privacy) grants researchers to use information entered to the Zooniverse platform (which includes its Talk pages) for the advancement of knowledge. Data exported from

Table 1: Overview of functionalities in *Talk* and the empirical indicators used in the present study.

	Threads	**#hashtags**	**@-messages**	**Hyperlinks**
Function	Organise subject matter	Sort and classify data	User interaction and notification	Referencing information
Empirical indicators	Length of threads, thread initiators (users).	Extraction of hashtags, frequency of use, clusters of users, and hashtags.	Extraction of @-messages, frequency of mentions, networks of users, and hashtags.	Extraction of http://-links, mentions of external resources.

the Talk forums contain no personal information (such as IP-addresses, cookies, e-mail addresses) other than the username selected by the users themselves. In this article, however, we mention usernames only in relation to researchers who have presented themselves on the Shakespeare's world "About" page. All other users are only presented with their respective roles, such as "moderators" or "super-users".

Result and Analysis

As a first step, we extracted and counted the frequency of use of the forum functionalities. A total of 11,450 posts were written by 388 individual users in 3,460 threads across 11 subforums ("boards"). Threads are often short: Only 10 threads have more than 30 posts, and the average thread is only 3.3 (median 2) posts per thread. 972 posts (8%) contain hyperlinks, 2,692 posts (24%) contain hashtags, and 2,483 posts (22%) contain the @-symbol used to *ping* (send a notification to) other users of the Shakespeare's world Talk Forum.[1]

RQ 1 – Distribution of roles

For our study, we distinguish three groups of users: Super-users (users contributing more than 100 posts to the forum each), active users (more than 10 posts), and casual users (10 or fewer posts). We arrived at the threshold of 100 posts for super-users as this was the approximate cut-off value for those users that had created 80% of the total amount of posts on the forum as a whole (including posts by very active moderators and researchers). Additionally, moderators and researchers in the project were identified and considered as separate roles. **Table 2** shows an overview of the roles and their quantitative contributions to the Shakespeare's World Talk forum.

A challenge in defining roles is that forums are dynamic and change over time. Moderators sometimes quit and new ones are assigned the task, and in Shakespeare's world there is even a case where one user became a researcher (however, this event is not in our current dataset). The exported data retrieved do not contain the metadata for which roles each user has. To reconstruct roles, we asked one of the principal investigators (Victoria van Hyning) to give us the dates (sometimes approximate) of when the roles were changed or assigned. With this information we were able differentiate posts made by moderators, researchers, and super-contributors, also marking up posts with the current status in cases where such roles have changed. While researcher and moderator are roles that are assigned as special qualities, our notion of super-users, active users, and casual users are quantitatively defined (see above).

The super-users produce 37% of the total forum content, thus, the contribution of these 11 users is substantially greater than the total of 363 other volunteer users. The second-largest group is moderators, although it should be noted that the 27% of forum posts by moderators are split up between only two individual users. Additionally, ten researchers contribute with 19% of the forum posts; four of these researchers have contributed with more than 100 posts each. Finally, the total number of contributions by 36 active users and 327 casual users account for only 10% and 7% of the forum contents, respectively.

These results indicate that the majority of the forum is created by a small number of users (17 users have created 80% of the posts). Among this group of productive users, we find the two most active moderators, four researchers, and 11 super-users. The distribution among these roles is, however, uneven. In total, the super users have created 4,226 posts among 11 users, and the four moderators have written 3,114 posts, with the most active moderator contributing 2,414 posts alone. Overall, this puts the two moderators at a much higher individual production in comparison to other roles.

Thread initiation

When the Talk-software was implemented, users were given the option of starting threads based on subjects encountered in the classification tasks instead of adding threads to a linear forum structure. This style of use is reflected in the Shakespeare's World forum, where 92% of the threads were initiated based on specific subjects. Another pattern in relation to thread initiation is visible in the user roles. **Table 3** shows the number of threads started by each group of users. Compared to the overall activity of the forum, the researcher-role stands out clearly: While the researchers contributed to about 19% of the overall forum posts, only about 1% of the threads in the forum are initiated by researchers. For the other user groups the distribution is more even. Moreover, the large bulk of subjects being transferred from the tran-

Table 2: Community roles in Shakespeare's World Talk (N = 11,450). While the number of posts takes into account the changing roles over time, this is not the case for the number of users. Here the moderator and super-user roles have changed over time, for example, one super-user became moderator, and two moderators quit over the course of the current dataset.

User role in forums	Criterion	Number of users	Number of posts	Percentage of posts	Posts per user
super-user	>100 posts	11	4,226	37%	384
active user	>10 posts	36	1,173	10%	33
casual user	<=10 posts	327	805	7%	2
moderator		4	3,114	27%	779
researcher		10	2,132	19%	213
TOTAL		**388**	**11,450**	**100%**	**30**

Table 3: Threads initiated by each user group (N = 11,450). Note the percentages for thread initiation are calculated in relation to the total sum of threads (N = 3,460) whereas the thread responses are calculated using the total forum posts which are not thread starts, but instead are responses (N = 7,990).

	Thread initiation		Thread responses		Threads initiated per person	Thread responses per person
moderator	1001	29%	2,113	26%	250	528
researcher	46	1%	2,086	26%	5	209
super-user (>100 posts)	1,413	41%	2,813	35%	128	256
active user (>10 posts)	551	16%	622	8%	15	17
casual user (<=10 posts)	449	13%	356	4%	1	1
Total	**3,460**	**100%**	**7,990**	**99%**	**9**	**21**

scriptions to the forum are initiated by super-users (41%) and moderators (29%), which reveals that these users are the ones driving the creation of new materials that are in the need or interest of being disseminated to the rest of the community. This way, it is possible to conclude that the researchers have a different role in knowledge formation. They rarely start new threads, but instead frequently respond to questions raised as the threads develop. While moderators and super-users also frequently respond, these user groups initiate a high number of threads as well.

This brings about another interesting finding. The casual users (sometimes referred to as the "long tail") have started a substantial amount of threads (449), but very few when broken down on each individual (in total 327 users). A similar pattern is found among the active users. They both have approximately a 1 to 1 ratio between thread initiations and thread responses, which indicates that they drop in and out quickly. Researchers, on the other hand, have the opposite behaviour. Starting up only 46 threads but writing 2,086 posts indicates that they act as expert advisors and make their contribution once the subject has been brought to their attention. Moderators and super-users instead bring out many new threads simultaneously as they write in the forum twice as much as initiating threads. This means that not only are they very active in bringing out new data for discussion, they also play an important and large role in discussing, disseminating, and analysing the data. Without the moderators and super-users, there would barely be an active knowledge production on the forum.

RQ 2 – Use of external knowledge resources

Understanding letters written in the 17th century requires contextual information, especially for verifying word meanings, checking historical dates and facts, and understanding which historical figures are present in the texts. Such knowledge lies outside the task of transcribing the documents. As we will show, however, this knowledge is valued highly by the volunteers and researchers alike. As the web contains numerous sources of information that are external to the Shakespeare's world forum, we have analysed the practice of bringing in external knowledge resources.

In this analysis we excluded the forum section "Help and Technical Issues" to better capture the knowledge practices that are directed to the research theme of Shakespeare's world. Firstly, we extracted all URLs by searching for the "http+ prefix. Secondly we expanded our searches by including indirect links (such as "I used Google to find …") (see https://github.com/christopherkullenberg/talk-analyzer). Then we coded them manually by the type of service used as an external resource, such as databases, archives, or social media.

Hyperlinking

Out of 10,605 forum posts, 885 (8%) contain one or more hyperlinks, marked by the character string "http." We investigated under which circumstances these were used and if the linking behaviour differed between user groups.

As shown in **Figure 1**, the different roles on the forum have different practices of bringing in information. Moderators mainly link to various project repository documents, i.e., documents up for transcription. Secondly, they refer to the project forum and websites, linking together various threads of discussion. These two major linking patterns indicate that moderators are indeed fulfilling their assigned task of bringing together and moderating the forum by shaping more coherent discussions. Researchers, on the other hand, are quite diversified in their linking practices. While they also perform the roles of the moderators to a certain degree, they frequently refer to the project forum and websites, comprising instructions, blogs, and other contextual information related to Shakespeare's world and the Zooniverse platform. This way, they anchor the citizen science contributions within their project, and show how the knowledge generated becomes valued and fed back into research. On example is an exchange between a researcher and a moderator. When transcribing, the moderator finds an occurrence of the word "esterdaie" seemingly used as a version of "yesterday" and tags the researcher responsible for the OED entries who responds "That's a great prompt for us at OED to do some more research. We'll note it carefully."

Often researchers thank the contributors by updating the website or writing a blog post about a discovery or phenomenon in the transcribed documents. Researchers are less active in referring to other forum posts, while they more often than moderators refer to historical repositories and databases such as Wikisource, Project Gutenberg, and Wellcome Library Documents, as well as dictionaries and thesauri. They often mention several sources in the same

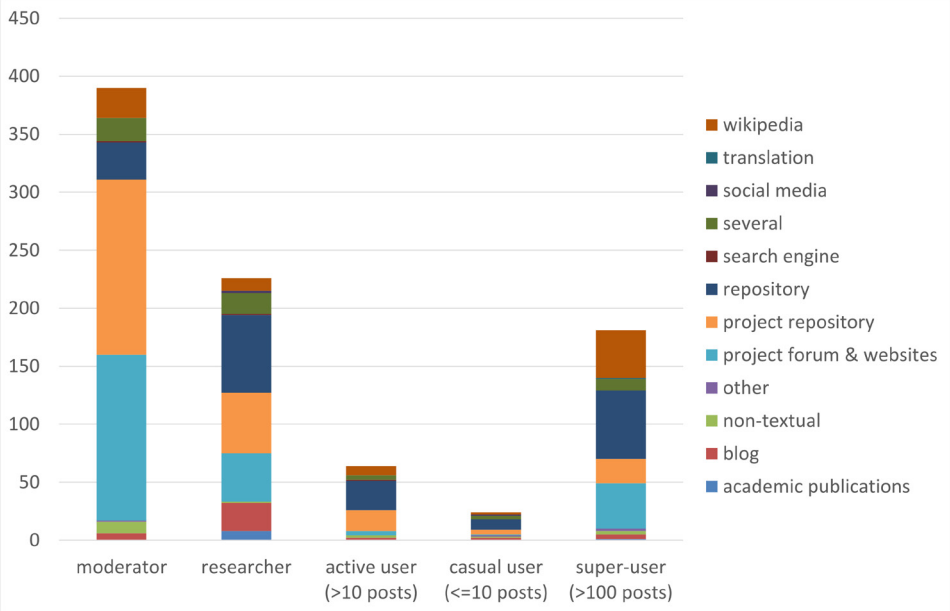

Figure 1: Use of internal and external hyperlinks containing the prefix string "http" (N = 885).

forum post, indicating a somewhat typical researcher style of writing, in which multiple sources are presented, and along the lines of knowledge synthesis, referring to an ongoing and cumulative research frontier. Sometimes these external resources are used by other roles once introduced as reliable sources by researchers. One such example is a link to commonly used symbols for measurement units in recipes (http://www.textcreationpartnership.org/docs/dox/medical.html), which was introduced by a researcher and then consequently used by super-contributors as a standard for interpreting such symbols. Furthermore, researchers also link to blogs, which they often have authored themselves, and to scholarly articles. Finally, the super-contributors are quite diversified in their linking practices. We find a more frequent use of Wikipedia, and the largest part of the links go to external repositories. This indicates that the super-contributors are wider in their quest for contextual information, going outside the project resources, often turning to free and open services such as Project Gutenberg, Wikipedia, and the Internet Archive.

Indirect linking

When users refer to online sources and search engines without hyperlinking (i.e., printing out the URL beginning with "http"), we call this indirect linking. It occurs more than 200 times in the forum material and is often expressed in the style of "I used Google," "I found on the Internet," or "I looked it up on Wikipedia." As this occurred frequently, we deemed it necessary to be the object of further analysis. However, the extraction of such phrases is not as straightforward as with conventional hyperlinks. We created a set of search queries based on the first retrieval of hyperlinks. For example, if users linked http://twitter.com or http://google.com, we generated search queries that would capture also indirect linking, for example "twitter" and "googl*" (for a complete list, see supplementary files at https://github.com/christopherkullenberg/talk-analyzer). This way we were able to exhaust most expressions, even though a limitation of this method is that we might miss ways of linking expressed by phrases unknown to us. However, while informal linking is rather imprecise, it does reveal a lot of interesting information about what forms of knowledge resources the users of the forum express as auxiliary resources. Such resources have a potential or actual capacity to bring in contextual information when transcribing and understanding the raw textual material.

The results of the informal links (**Figure 2**) show firstly that they are used by researchers the most, followed by the super-users. However, researchers tend to link indirectly back to the project forum and its websites (especially the Shakespeare's World blog), to various repositories (especially the OED), and to social media, while super-users most often refer to search engines, a behaviour shared with moderators. Moreover, moderators, active users, casual users, and super-users often refer to Wikipedia in their posts, while this is less common in the linking practices of researchers. This suggests that researchers are less willing to refer indirectly to free and open search engines and Wikipedia, while this appears to be favourable by all other users. Also, when breaking down the repositories, we find a stark contrast between researchers and super-users. While researchers point to the OED, the super-users more often refer to open repositories, such as the Dictionary of Scots Language (http://www.dsl.ac.uk/). We often find the expressions "on Google," "the dictionary," or even "on the internet," which are imprecise and sometimes impossible to know exactly what is referred to (which website, which dictionary, etc.).

These indirect links are interesting to look at in a detail, because web searches can be used for almost any query

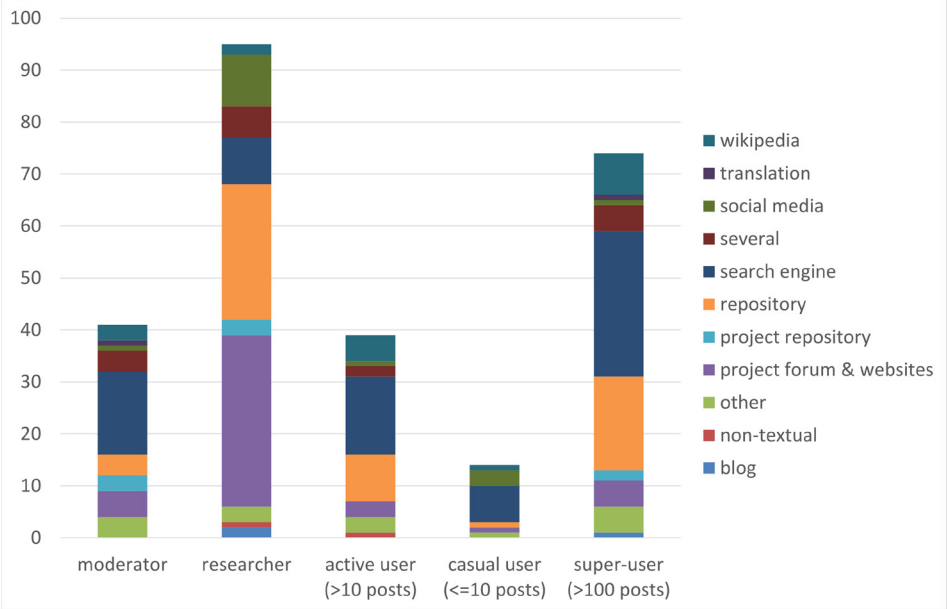

Figure 2: Use of indirect links (N = 263, see supplementary files at https://github.com/christopherkullenberg/talk-analyzer for a full list of search terms).

and appear to be used when the user is not sure where to look or how to exactly phrase the question. Notable examples among super-users and moderators include:

> "The internet says this is Esztergom in Hungary. Its medieval Latin name was *Strigonium* and it has a history of battles against Turkey." (moderator)
> "A quick Google tells us that the Beaumont family were indeed involved in coal mining so a foray into Welsh mining sounds reasonable – if unprofitable!" (moderator)
> "I Googl'd the word ayenst and found that it occurs in the Canterbury tales. From the context, it seems to have the same meaning as "against"" (super-user)
> "So ... Anyone know what types of cows were around in Shakespeare's time? The "red" ones seem to have been singled out for this recipe implying that there were others. I could Google. In fact I might.)" (moderator)

The overall practice suggests that searches are used to find contextual information that makes the transcribed text more meaningful. As in the examples above, users wish to find a historical place, a historical family, or the meaning of an old word not found in the average dictionary. This is also evident in the super-users' frequent references to historical sources and documents that are brought in as points of reference.

RQ 3 – Use of hashtags

Using hashtags is a common affordance in social media and has been implemented in the Shakespeare's world forum. In this section we are interested in how hashtags are used to systematise knowledge, and if the practice differs between user groups.

Hashtags primarily used by researchers, moderators, and super-users

The hashtags that are predominantly used by researchers, moderators, and super-users are: #catholic, #OED, #paper, #womanwriter.[2] Out of these tags, #catholic, #OED, and #womanwriter were created as sub-forums from the beginning of the project and taken up by users as the project moved on. These all have in common that almost all tags are created by these categories of users. Often the super-users are the ones creating the most hashtags in quantitative terms.

In **Figure 3** we see the #catholic(s) hashtag as it is played out in our dataset over time. The first use is made by a researcher (Victoria van Hyning), but is almost instantly picked up by a handful of super-users, and a bit later by moderators. Most of the #catholic(s) post are made in the first six months of the project (this is also the case for the total data produced). However, in mid 2017 there is a spike in researcher contributions again, and when looking closer to these posts we see that the researcher is tagging up older threads by replying to users who have found interesting texts mentioning catholic writers.

The #OED hashtag follows a different trajectory. It is heavily used in its first six months by researchers, who are soon aided by moderators and super-contributors. However, a year later it is once again picked up by another constellation of users, consisting mainly of active users, super-contributors, and moderators. The #OED hashtag is of special importance because it is connected to one of the main goals of the project, namely to collect new words for the OED. This way we often observe users tagging their posts with #OED as they expect an expert or researcher to come into the thread and confirm or reject the word as a new possible entry to the dictionary. As most users outside university libraries do not have access to the paid service of the OED, they are unable to confirm their finding

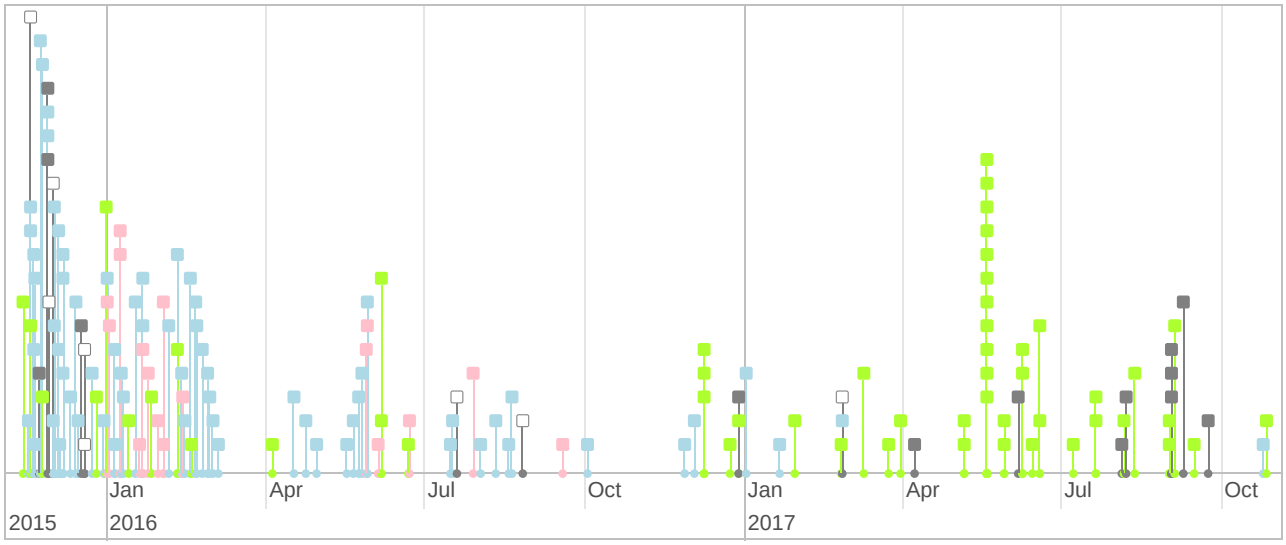

Figure 3: Timeline for the use of the #catholic(s) hashtag. Green = Researcher; Pink = Moderator, Lightblue = Super-user; Grey = Active user; White = Casual user. The code for generating this figure is available at https://github.com/christopherkullenberg/talk-analyzer.

on their own, and thus turn to the forum. As we saw in the previous section on external resources, we found numerous links to open dictionaries, indicating that some users also try out to find out the meaning of newly found words.

The #paper hashtag has an interesting dynamic of interaction. It was initiated by a super-user (who later became moderator) noting that one author used different qualities of paper for different letters, reserving expensive and fine paper only for important messages. The same hashtag was later used by researchers, in particular one called Elaine Leong who, a few days later, announced, "I'm doing a project on paper in recipes." As this was written in the forum, many more users begun directing their messages towards her username, @elaineleong (see RQ 4), as soon as they found examples of paper in historical recipes in their transcriptions. So, we can detect an interesting form of collective knowledge production between researchers, super-users, and moderators, which is interwoven with the aid of a hashtag, and becomes a popular one used by researchers and users alike, even though it was not used as a subforum from the start of the project.

The hashtag #womanwriter(s) was first mentioned in a post by a moderator, but was then frequently used by super-users in the first months of the project. In a similar style as with the #catholic(s) hashtag, the researcher Victoria van Hyning makes frequent use of the hashtag more than a year later, in summing up the material collected. This way the data collected in the first phase becomes accessible to research as the hashtags then can be used to find a collection of materials relating to a particular theme.

Volunteer-driven hashtags

Volunteer driven hashtags include #medical, #medicine #recipe, #bleedthrough, #latin, #cooking, #food, #letter. These have in common that researchers produce less than 10% of the tags. Instead it is mainly the super-users that produce #letter(s) and #bleedthrough. For the other tags, the main contribution stems from super-users and active users. None of these hashtags have their own subforum, but have instead grown dynamically from the heavy use of them by non-researchers. In **Figure 4** we see the #medical hashtag to which researchers only make four contributions, while the bulk of the tagging is performed by super-users and active users. None of the researchers indicate that they are specifically doing research on medicine in Shakespeare's time. Still, the hashtags #medicine and #medical continue to be widely used by non-researchers, indicating that this is an important topic of interest to them.

As a conclusion, we see the volunteer-driven hashtags having in common that they describe phenomena unforeseen by the researchers, which instead are taken up by volunteers. There are no sub-forums for these hashtags, which suggests that the frequent texts on medicine were not expected or deemed interesting by the project creators. They are, however, of great interest to the volunteers, who invent and use these tags frequently.

RQ 4 – Use of @-messages

The Talk forum software used in Shakespeare's World allows so-called "pings" (@-messages) to alert users about discussions. By writing the @-symbol in front of a user name, discussants can insure that this user will be alerted as they log in. Although this feature is used to some extent by all user groups, the most active users of pings are the researchers on the platform, who were the senders of 1,998 and receivers of 1,413 pings.

As shown in **Table 4**, moderators more often send pings than receive them. This was somewhat surprising, because we expected moderators to often be asked to answer questions. While this is sometimes the case, moderators are also very active users, and include in their moderator role to ping users who might know more about something or thank users who have contributed. Another surprise is the distribution among researchers. Here we also find a higher frequency of sending pings than receiving them. Partly this is because at least one

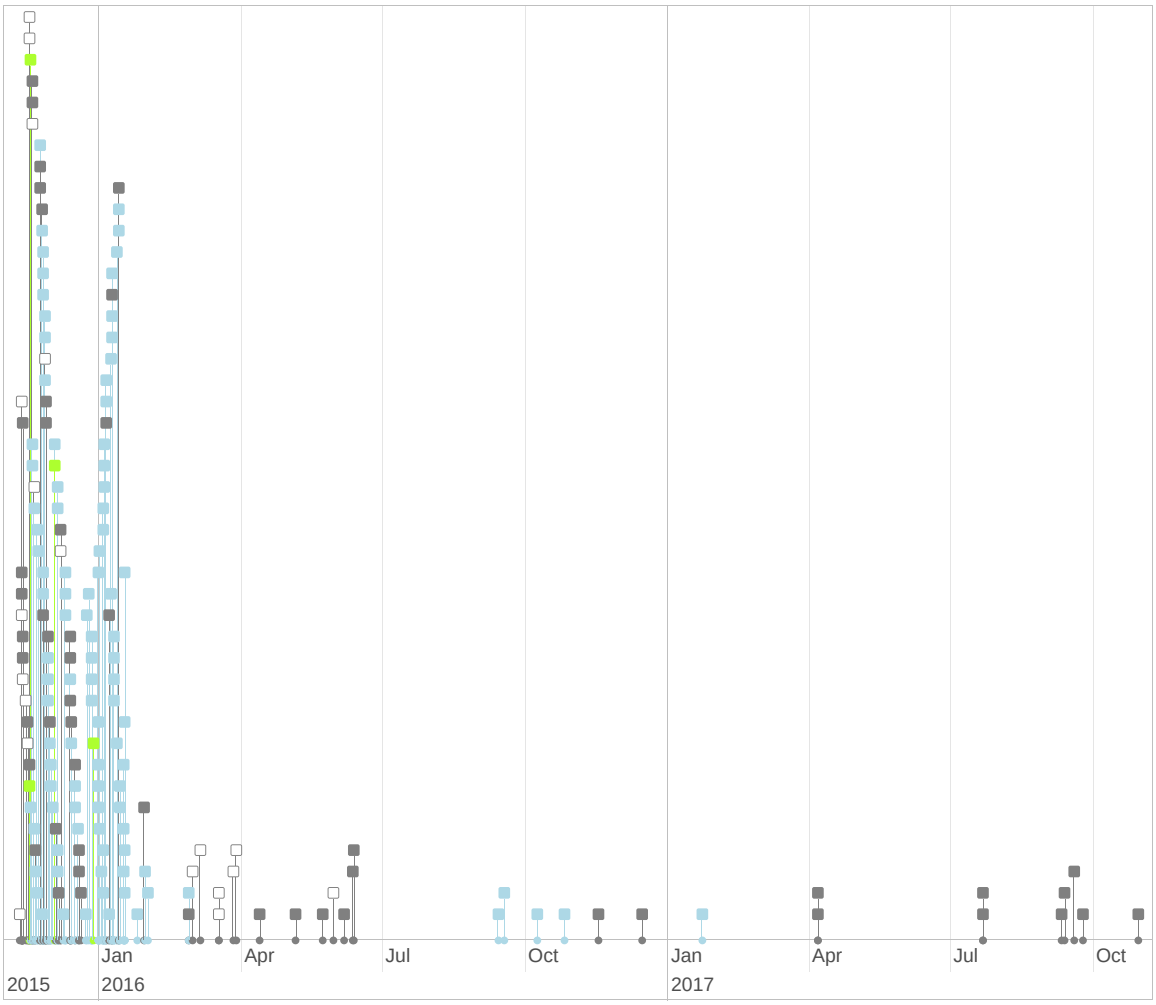

Figure 4: Timeline for the use of the #medical hashtag. Green = Researcher; Pink = Moderator; Lightblue = Super-user; Grey = Active user; White = Casual user. The code for generating this figure is available at https://github.com/christopherkullenberg/talk-analyzer.

Table 4: Pings (@-messages) sent and received by user groups N = 3632.

	Pings sent	Pings received
moderators	992	643
researchers	1,998	1,413
super-users	471	666
active and casual users	171	910
Total	3,632	3,632

researcher behaves almost like a moderator, pinging users that might know more, or, as is often the case with all researchers, they frequently thank the users when they have contributed. Super-users receive more pings than they send, and this is also the case for active and casual users. Especially in the latter case, the large number of active and casual users are receiving individual thank-you messages and answers to questions.

If we extract both hashtags and @-messages from each post, we are able to study the co-occurrences of hashtags and @-messages from a network perspective. Here we took each post containing at least one hashtag and one @-message and created a directed network from the hashtag to the pinged user. This way we are able to visualize a cluster, which suggests a particular distribution of topics.

In **Figure 5** we can see how the #OED hashtag appears as a center of gravity, with the majority of these messages being directed towards @PhilipDurkin. Philip Durkin is a researcher working at the OED, hence this is unsurprising. However, many other moderators and researchers also receive #OED-tagged messages, indicating that the discovery of new words is not confined to a single gatekeeper, but attracts a wider community of users and discussions. Almost all researchers are pinged in connection with the #OED hashtag, suggesting that this hashtag serves as an important communication point between researchers and non-researchers. However, the two active moderators are also being pinged in frequently, equally often as active researchers.

Another type of cluster is formed between the researcher Victoria van Hyning (@vvh), who has frequent pings with the two active moderators, especially on the hashtags #catholic and #womanwriter, but also on several more. In other words, here we find an even more genuine example of researcher-volunteer knowledge exchange. One

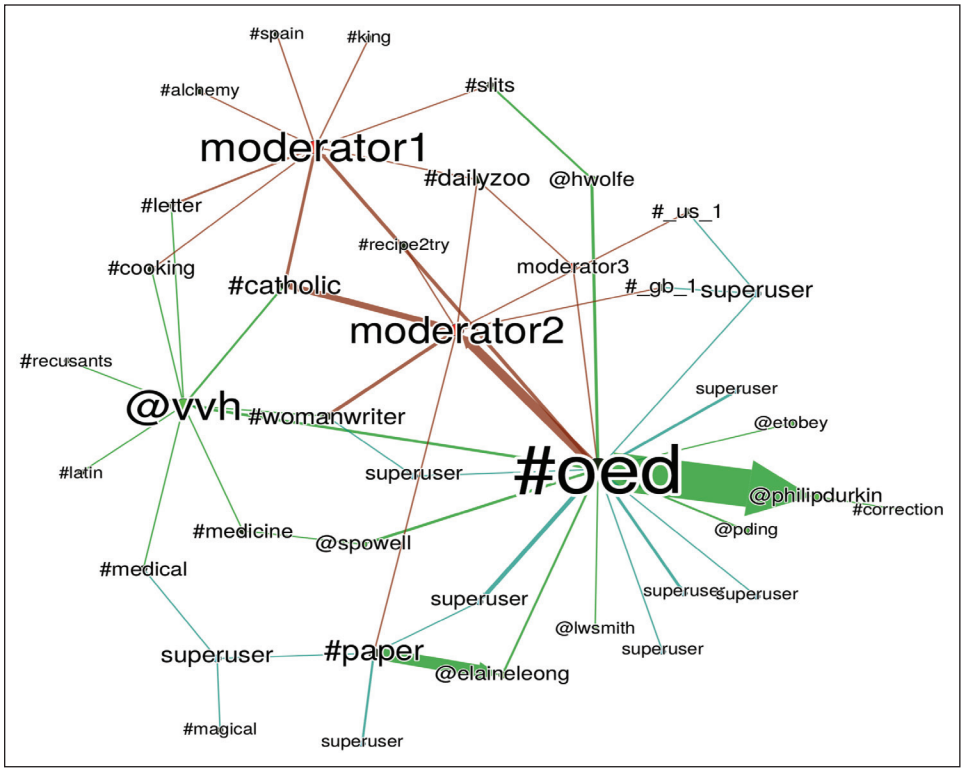

Figure 5: Indegree network – hashtag to username. The data were extracted by matching posts in which there were both hashtags and pings. We excluded such co-occurrences if they happened only once in order to reduce data (this effectively excluded contributors below super-users and disconnected nodes in the network). Moreover, the Force Atlas 2 algorithm (Blondel et al. 2008) only concentrates nodes that are connected more than two times to the center of gravity to which the figure was limited. Legend: Green = researcher; red = moderator; magenta = superuser; black = hashtag. The code for generating this figure is available at https://github.com/christopherkullenberg/talk-analyzer.

striking example of combining hashtags and @-messages is the following post by an active user:

> @VVH #recusants #catholics
> Interesting stuff about crosses and badges of idolatry but I also note that this letter was dated 20 July 1588. According to Wikipedia, the following day the English fleet attacked the Spanish off Plymouth. The fight continued into September. This letter is fairly permeated with apprehension, for good reason, I'd say: enemies abounded, both external and internal.

Here we can see how Victoria van Hyning is pinged and the hashtags relevant to her research interests are used in the post. Moreover, the external resource Wikipedia was used to contextualize the finding, and finally the findings are interpreted.

Another case of where this takes place is on the #paper hashtag, where Elaine Leong (@elaineleong) is pinged according to her research interest of paper-use in recipes. Finally, there is a case of a moderator-only hashtag, #dailyzoo, in which moderators ping each other exclusively.

In total, we conclude that researchers are the most commonly pinged users, followed by moderators and super-users. In many instances, researchers are pinged because they are asked for their expertise, but such expertise has sometimes been achieved also by moderators and super-users. Among the 15 most pinged users, 7 of the total 10 researchers in the team appear. Moreover, because moderators and super-users are very active in transcribing texts and bringing these to the forum, they are often thanked by the researchers who show their appreciation for their volunteer work by pinging them.

Summary of findings

One of the characteristics of online environments is that they allow organizing information through tagging, pinging, and linking. **Table 5** shows an overview of these features in Shakespeare's World Talk, indicating that the different user groups make use of them in different ways.

This reveals another dimension of a citizen science project that primarily has the purpose of transcribing texts. The different uses of the forum means that the users attain different roles fulfilling different functions with regards to knowledge production. On the most general level, we find different patterns of online behaviour among researchers, moderators, super-users, and other users. The moderators and super-users bring out new texts for discussion, as they are the ones doing most of the transcription work that ends up as issues on the forum. The researchers generally only respond once an issue has been posted, and then they often generously provide contextual information and means of understanding the texts. They also tag the threads and thank the volunteers for their contribution,

Table 5: Summary of the non-linear forum organization by user groups (N = 11,450) and research questions. The percentages are calculated in relation to the number of posts per user group.

User roles (RQ 1)	Number of posts	Tagging (RQ 3)	Pinging @ (RQ 4)	Linking http (RQ 2)	Indirect linking (RQ 2)
moderator	3,114	351 (11%)	822 (26%)	410 (13%)	53 (2%)
researcher	2,132	254 (12%)	1,156 (54%)	255 (12%)	103 (5%)
super-user (>100 posts)	4,226	1,542 (36%)	371 (9%)	186 (4%)	78 (2%)
active user (>10 posts)	1,173	410 (35%)	94 (8%)	89 (8%)	50 (4%)
casual user (<=10 posts)	805	135 (17%)	40 (5%)	32 (4%)	21 (3%)
total	11,450	2,692 (24%)	2,483 (22%)	972 (8%)	305 (3%)

and they often are pinged into a conversation, as many users expect more information about the sometimes difficult challenge of understanding a historical manuscript.

Discussion

Moderation is an important task to any forum. In our study we have shown that the designated moderators are investing a lot of time in tagging and pinging the discussion in order to bring the right person to the right issue. Also, some researchers may display this sort of moderator behaviour, as they know the project very well and link to resources or other researchers to expand a discussion. We distinguished three somewhat different linking behaviours among moderators, researchers, and super-users. While researchers have a quite typical academic way of making references to sources outside the project, linking to research and other repositories, the use of open sources such as Wikipedia is more common among super-users and active users. Also, when there is limited access to a resource, for example the OED, users often ping a researcher with proper access.

We found that the use of hashtags was an important feature of the forum, especially for improving the search functionality and thus systematising the forum information. However, because anyone can write any hashtag without restrictions (just like on Twitter or Instagram), the issue of whether to control hashtags appears. If not controlled, there is a risk of overlap, for example between #medicine/#medical, #recipe/#cooking/#food, etc. However, one of the seeded tags that was introduced by researchers, #Recipes2Try, was not picked up by the users as much, perhaps because it is hard to type, or because most recipes are not recommended to try at all. Instead the volunteers used #recipe, #cooking, and #food to further differentiate the findings, and perhaps making it easier to tag the content.

Conclusions

Summing up, we propose that hashtags and pings afford an interesting mode of organization of the knowledge produced in Shakespeare's world. The free use of hashtags enables the emergence of new contrasts in the disseminated material, where new phenomena and discoveries can be made. The ping function, in turn, enables the formation of sub-communities of interest along certain topics, which concentrates the issues discussed and allows for special expertise to be developed and allocated at the right spot. It also creates a direct link between researchers and volunteers, which has led to some interesting collaborations in knowledge creation outside the regular constraints of academic humanities research.

Based on these conclusions, we suggest that VCS projects consider the following issues with regards to collaborative forums and knowledge sharing spaces. Firstly, we have shown that forums play an important role for understanding and interpreting the designated task (in our case transcription). These interaction spaces do more than just motivate people to do more citizen science; they also provide an interface between data and knowledge, where volunteer contributors can both create and immerse themselves in knowledge practices. This is an important finding because it shows that the discussion forum plays an important role in knowledge creation as such, not only as a motivational driver for transcribing more texts. We also have seen how new phenomena are being discovered by volunteers, something which should be valued highly and credited properly. This is similar to Luczak-Roesch et al.'s (2014) conclusion that citizen scientists can contribute to "coordination around hypothesis" and "serendipitous citizen-led discoveries" when using features of the Talk forum. However, unlike Kasperowski and Hillman (2018), we have not found instances of tensions among user groups in our material. This could be due to the smaller size of the forum and younger age of the project, but also might hint at a difference between the domains of natural sciences in Kasperowski and Hillman's study as opposed to the humanities in our study. Secondly, we also have shown the importance for researchers to respond and share their expertise with volunteers, and researchers who do that are rewarded with highly motivated citizens as co-researchers in their projects.

Our study is limited empirically to only one citizen science project, Shakespeare's World. Because citizen science is such a diverse phenomenon, patterns of interaction between user groups may differ depending on what the project is about, the number of users, and the engagement of researchers and moderators in the forums. Another limitation in our research design consists of giving more weight and attention to active and frequent users of the forum, even though they represent only a small number of the overall user base. Less active users or users who only read but do not write on the forums (so-called lurkers)

leave fewer or no "traces" that can be analyzed with trace ethnography.

Forums in VCS play an important role in co-creation of knowledge. Thus, they must be regarded as having more than a motivational or supporting function. We have shown that new knowledge is produced as volunteers and scientists collaborate on VCS forums. For project owners and researchers who want to conduct citizen science, there is an important lesson to learn from such a conclusion, namely: Forums must be carefully maintained, which means that researchers have to be present on the forums and be prepared to reciprocally exchange knowledge. This means that researchers should help, encourage, and share their knowledge with volunteers. In return, they are rewarded with active collaborators beyond the traditional academic setting.

Notes

[1] The exact numbers of @-messages, hashtags, and URLs vary between Regular Expressions engines. In this article we use the engine built into Microsoft Excel 2016, which we verified against Python3's regexp engine. The results differed only in the range of 5–10 hits.

[2] This hashtag is also expressed as #womenwriter(s) by 5 users with the same meaning. On one occurrence this form predates #womanwriter, in a post by Victoria van Hyning, who could be called the initiator of the hashtag as a whole.

Acknowledgements

The authors wish to thank Dr. Victoria van Hyning for providing the forum data for the Shakespeare's world forum and for generously sharing contextual information about how the project has developed since it was launched.

Competing Interests

The authors have no competing interests to declare.

References

Bastian, M, Heymann, S, Jacomy, M and **others.** 2009. *Gephi: an open source software for exploring and manipulating networks*.

Blondel, VD, Guillaume, J-L, Lambiotte, R and **Lefebvre, E.** 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008. DOI: https://doi.org/10.1088/1742-5468/2008/10/P10008

Geiger, RS and **Ribes, D.** 2011. Trace ethnography: Following coordination through documentary practices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference* on (pp. 1–10). IEEE. DOI: https://doi.org/10.1109/HICSS.2011.455

Golder, SA and **Huberman, BA.** 2006. Usage patterns of collaborative tagging systems, *Journal of Information Science*, 32(2): 198–208. DOI: https://doi.org/10.1177/0165551506062337

Hedges, M and **Dunn, S.** 2017. Google-Books-ID: SNvWDgAAQBAJ. *Academic Crowdsourcing in the Humanities: Crowds, Communities and Co-production*. Chandos Publishing.

Kasperowski, D and **Hillman, T.** 2018. The epistemic culture in an online citizen science project: Programs, antiprograms and epistemic subjects. *Social Studies of Science*, 48(4): 564–588. DOI: https://doi.org/10.1177/0306312718778806

Letierce, J, Passant, A, Breslin, J and **Decker, S.** 2010. *Understanding how Twitter is used to spread scientific messages*.

Liberatore, A, Bowkett, E, MacLeod, CJ, Spurr, E and **Longnecker, N.** 2018. Social Media as a Platform for a Citizen Science Community of Practice. *Citizen Science: Theory and Practice*, 3(1). DOI: https://doi.org/10.5334/cstp.108

Luczak-Roesch, M, Tinati, R, Simperl, E, Kleek, MV, Shadbolt, N and **Simpson, R.** 2014. *Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science*. In: 4 June 2014 United States. p. 10.

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

Reed, J, Raddick, MJ, Lardner, A and **Carney, K.** 2013. An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects. In: *2013 46th Hawaii International Conference on System Sciences*, 610–619. January 2013. DOI: https://doi.org/10.1109/HICSS.2013.85

Sinclair, J and **Cardew-Hall, M.** 2008. The folksonomy tag cloud: when is it useful? *Journal of Information Science*, 34(1): 15–29. DOI: https://doi.org/10.1177/0165551506078083

Tinati, R, Van Kleek, M, Simperl, E, Luczak-Rösch, M, Simpson, R and **Shadbolt, N.** 2015. Designing for Citizen Data Analysis: A Cross-Sectional Case Study of a Multi-Domain Citizen Science Platform. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 4069–4078. CHI '15. 2015 New York, NY, USA: ACM. DOI: https://doi.org/10.1145/2702123.2702420

Trant, J. 2009. Studying Social Tagging and Folksonomy: A Review and Framework. *Journal of Digital Information*, 10(1).

Trant, J, Bearman, D and **Chun, S.** 2007. The eye of the beholder: steve.museum and social tagging of museum collections. In: *Proceedings of International Cultural Heritage Informatics Meeting (ICHIM07)*. p. 2007.

Veletsianos, G. 2017. Three Cases of Hashtags Used as Learning and Professional Development Environments. *TechTrends*, 61(3): 284–292. DOI: https://doi.org/10.1007/s11528-016-0143-3

Zhang, A, Zheng, M and **Pang, B.** 2018. Structural diversity effect on hashtag adoption in Twitter, *Physica A: Statistical Mechanics and its Applications*. 493: 267–275. DOI: https://doi.org/10.1016/j.physa.2017.09.075

How to cite this article: Rohden, F, Kullenberg, C, Hagen, N and Kasperowski, D. 2019. Tagging, Pinging and Linking – User Roles in Virtual Citizen Science Forums. *Citizen Science: Theory and Practice*, 4(1): 19, pp. 1–13. DOI: https://doi.org/10.5334/cstp.181

Submitted: 25 May 2018 **Accepted:** 04 February 2019 **Published:** 07 June 2019

]u[*Citizen Science: Theory and Practice* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS