End-user programming on large online knowledge sharing platforms as collective epistemic resource producing activity

Thomas Hillman, Alena Seredko
thomas.hillman@gu.se, alena.seredko@gu.se
University of Gothenburg

Abstract: Large online knowledge sharing platforms such as Stack Overflow, the platform in focus in this study, are sites of learning for millions of people. As users share knowledge through these platforms, they generally interact with Graphical User Interfaces (GUIs) that are provided by platform owners. In some cases, however, users with programming skills write scripts that allow them to make use of platform data in ways beyond those provided in standard GUIs. The results of our study of scripts produced for Stack Overflow show how scripting can be a collective activity that results in new epistemic resources. These epistemic resources suggest a widespread interest in revealing and unpacking the ranking mechanics on the platform and in understanding how to contribute most effectively in terms of those rankings.

Online platforms such as discussion forums, video sharing sites and Community Question Answering (CQA) sites are major venues for learning and knowledge sharing. As each successive generation of platforms develops, they offer new ways of producing and interacting with knowledge that users exploit and reconfigure in an ever-evolving cycle of change. While there is a great variation in ways of producing and interacting with knowledge, common to online platforms is that the digital traces left by users are the primary way through which users, including researchers, and systems makes sense of and produce value from knowledge sharing practices (Geiger & Ribes, 2011). Users generally produce and interact with the digital traces that constitute the platforms they use through Graphical User Interfaces (GUI) provided by platform providers such as websites or mobile apps. In some cases, however, users create other forms of interface that allow programmatic interaction with digital traces through an Application Programming Interface (API), scripting interface or techniques such as web-crawling. These forms of interface allow access to the platform databases that exist behind the scenes of a commonly used GUI, produce new datasets that reconstruct the information contained in those databases, and allow programs to be written that interact with platforms on behalf of users.

In this paper, we argue that the scripts users write to make use of the trace data on online platforms in alternative ways to those provided in the GUIs offered by platform providers constitute epistemic resources that may be collectively produced and used in the sense that they are shared ways of making meaning from platform data. To examine this phenomenon, we investigate the scripts written by users of the programming focused CQA platform Stack Overflow and the interactions between users that are associated with the creation and use of those scripts. This examination is guided by two research questions:

- How is scripting on knowledge sharing platforms a collective activity?
- What epistemic resources do end-users produce through scripts on knowledge sharing platforms?

Online knowledge communities and collaboration at scale

Within the Computer Supported Collaborative Learning (CSCL) community there is a long tradition of examining knowledge sharing practices in relation to digital tools (van Aalst, 2009). Much of this research has conceptualized collaboration in relation to interactions between individuals or within relatively small groups often in terms of Communities of Practice (Lave & Wenger, 1991). Some work has considered larger knowledge communities in formal educational settings by contrasting the details of discourse patterns with those in smaller communities (Fu et al., 2016), while other work has offered new conceptualizations for larger knowledge communities such as Communities of Inquiry and Cultures of Participation (Fischer, 2016). However, there have also been several calls for more consideration of what might constitute knowledge practices and collaboration in the kinds of very large and often ephemeral groupings that can be found on digital platforms such as social media (Wise & Schwarz, 2017). A key response to such calls is the work of Jeong et al. (2017) where the authors propose a framework for distinguishing different types of interaction in online knowledge communities. The distinctions between these types draw on earlier distinctions made in relation to technologically mediated interactions in smaller groups (Dillenbourg, 1999), but also importantly incorporate the idea of *stigmergic* interaction that was originally used to describe the activity of insects such as a collective of ants building an anthill. Following this line, large scale online communities can be seen to function through stigmergic interaction with the affordances of the technological

platform they are based in providing opportunities for collective activity. Considering stigmergic interaction with other more direct forms of interaction, Jeong et al. (2017) propose a continuum with four types of collective activity moving from *attendance*, "where people behave according to their individual goals" through *coordination*, *cooperation* and finally *collaboration* where "people behave according to shared goals" (p.140). The question of what types of interaction take place through a particular digital platform is an empirical one that must consider both the affordances of the platform and the practices of those who use it.

For the purposes of this study, the platform in focus, Stack Overflow, is understood to facilitate a knowledge community in the sense that members jointly engage in knowledge-related practices (Paavola et al., 2004). The platform is popular for both professional and hobby software developers or programmers. Site owners, Stack Exchange, promote the site as "Where developers learn, share, and build knowledge" (Stack Exchange, n.d.-b) and describe it as having more than 50 million visitors a month. Unlike other popular knowledge sharing platforms such as the discussion forums on Reddit and the peer-produced encyclopedia Wikipedia, Stack Overflow is structured as a Community Question Answering (CQA) site. In this way, Stack Overflow is like earlier platforms such as Yahoo! Answers and Answer Garden in that interactions are generally constrained to the form of question-and-answer sequences. These sequences or threads might include several answers to the same question and serve as knowledge repositories on the question topic.

Much like other CQA sites and knowledge sharing platforms with different structures like Reddit, Stack Overflow involves several platform mechanics such as voting and badges that provide additional traces. These mechanics are used to produce metrics that value and rank user practices such as the quality of questions and answers posted. The primary metric on the platform is called 'reputation score', calculated by combining the influence of different traces such as votes received on user's posts. It is described by the platform owners as "a rough measurement of how much the community trusts you; it is earned by convincing your peers that you know what you're talking about" (Stack Exchange, n.d.-c). Reputation score produces the overall ranking of users and is complimented by various badges awarded for performing practices. These metrics are then presented on users' profile pages, when users post questions, answers, and comments, and on pages where rankings are displayed. In this way the metrics are of concern to many members and, as we demonstrate in the results, they are often the focus of the scripting activity that we examine in this study, a practice that can be understood as a form of end-user programming (Nardi, 1993).

End-user programming

In contrast to end-user development where design processes are opened to users' innovations (Fischer et al., 2004), the focus of this study is end-user programming where users with the necessary competencies and agency program modifications or extensions to existing software (Nardi, 1993). The form of end-user programming that is in focus is scripts written to interact with a feature offered by Stack Overflow owner's, Stack Exchange, called Stack Exchange Data Explorer (SEDE). This feature uses an example of what Nardi (1993) describes as a task-specific programming language that allows users to retrieve data from the databases associated with any of the Stack Exchange community sites. It is particularly well used in relation to Stack Overflow where many of the professional and hobby programmer members have the necessary competencies to make use of it. The data accessible through SEDE is available through a Creative Commons (CC BY-SA) license where reuse is freely allowed so long as attribution is provided, and the resulting work is also shared on similar terms. As Stack Exchange founder Jeff Atwood describes in a blog post,

> "This means it belongs to *everyone*, and can be freely reused (even commercially!), so long as it is [*sic*] follows our simple rules of attribution. That's our contract with the community — it's *your* generously contributed content that makes these websites worth visiting in the first place!" (Atwood, 2010)

This conception of community data ownership means that Stack Exchange datasets are widely and even unusually available when compared to other online platforms. They are available through services such as Google's BigQuery data warehouse and through an Application Programming Interface (API), but the SEDE interface is particularly accessible since it can be found on the Stack Exchange website, does not require additional programming environments or large downloads, and relies on common Structured Query Language (SQL) queries.

All queries that are composed on SEDE are stored and made available to other users. Each time they are used, they retrieve, operate on and present different aspects of the activity on the platform in real-time. In this sense and similar to arguments made by Fischer et al. (2004) in relation to end-user development generally and Wahl et al. (2017) in relation to query based information retrieval, scripts can be understood as epistemic resources. At the time of writing, there were nearly 2 million queries posted. Users can compose new queries or search for and use queries that were written by others, and favorite those they want to return to. They can also 'fork' existing queries allowing them to create independent versions that can be modified while maintaining their lineage. Since

SEDE runs on Microsoft servers, queries are composed using the syntax of the Transact-SQL (T-SQL) variant of SQL. The data that can be accessed through queries are also well-defined with a database schema that articulates an ontology of available data (Stack Exchange, n.d.-a). This means that queries follow a regular syntax and operate on well-defined data objects making them relatively straight forward to evaluate and modify. Users also have the option of giving their query a title and providing a description of its function. In this study, we make use of the readability of the queries along with the titles and descriptions provided by users to understand the epistemic resources that they constitute.

Method

The general approach taken in this study draws on trace ethnography (Geiger & Ribes, 2011) in the sense that the digital traces users leave on Stack Overflow are understood to be the primary means through which they interact with each other and the platform, and are thus a means by which we as analysts can come to know their activity. The trace data collected forms two different corpora, each allowing for analysis that addresses one or both research questions. For the question of how creating alternative interfaces to Stack Overflow data through query writing is a collective activity, a corpus of interactions between users about queries and query writing was assembled. This was then complemented with a corpus of queries themselves that was also examined in relation to the question of what epistemic resources scripts constitute. Both corpora were collected on May 27th, 2021 and their collection relied on access to publicly available data through the Stack Exchange Application Programming Interface (API) or through web crawling techniques. While all data on the Stack Exchange platform are made available for reuse based on a Creative Commons license that only requires appropriate credit be given, concerns related to the integrity of community members when those data are reused for research purposes mean that there are ethical concerns above and beyond concerns associated with access or ownership of data. In response, a situated approach to research ethics is taken in this study where consideration is paid to how details of the content contributed by individuals in the assembled corpora are shared.

Assembling the corpora

Stack Exchange has extensive opportunities for interaction and for each main question and answer site such as Stack Overflow, there is a 'meta' site where users can ask questions about the site. There is also the opportunity to create and participate in persistent chatrooms associated with each site where conversation can take place without following the strictly regulated question and answer format of the main and meta sites. Conversations on the topic of SEDE and writing queries for use with it in relation to Stack Overflow primarily take place in two chatrooms, one associated with the main site and the other associated with the meta site. For this study, the interactions in both these public rooms were collected to form a single corpus. This was achieved through a *Python* script that requested data from the Stack Exchange API and generated a table with the content of each post, the user who posted it, the time it was posted at, the room it was posted to, and if applicable which post it replied to.

Unlike the sites and chats hosted by Stack Exchange, there is no API with which to access SEDE queries. Without access to an API, a web crawler was written that automates the process of visiting the page for each SEDE query and collecting the relevant information. This spider was written in *Python* using the *Scrapy* library. It visited each Stack Overflow SEDE query page and collected the text of the query itself along with the title, description, creator, creation date, number of views, and the number of stars assigned to it by other users. This resulted in a table with the body and meta data for 26,370 queries.

Analyzing the corpora

The analytic approach taken follows the argument made by (Lindgren, 2020) amongst others that the scale and structures of trace data produced by digital platforms necessitates the use of computational methods, but that an interpretivist stance can be maintained when making sense of them. Both the corpus of chat interactions about SEDE and the corpus of SEDE queries themselves were examined in relation to the question of how scripting on knowledge sharing platforms by end-users is a collective activity. This analysis proceeded from the argument made by Jeong et al. (2017) that the scope of what constitutes coordination or even collaboration through asynchronous online platforms may be widened if stigmergic along with more direct forms of interaction are considered. First, evidence for direct forms of interaction was examined in the corpus assembled from the two chat rooms on the topic of SEDE script writing and use. Basic descriptive statistics such as the number of discussion threads, individual posts and replies were compiled before a table matching posts and replies with the associated user ids was created using the *Pandas* library for the *Python* programming language. This table was then used to produce a network of replies where reciprocal interactions between users could be graphed using the network analysis program *Gephi*. Second, evidence for indirect forms of interaction was examined in the corpus of SEDE queries. This involved examining the number of 'forks' of each query, where a 'fork' is the act of taking

existing code, in this case a query, and making a copy of it so that it can be developed independently. On SEDE, the action of 'forking' a query is supported by a button that automatically generates a complete copied instance of an existing query. Since this action is not recorded as a data item that is accessible through SEDE, the presence of multiple queries with the same name or description was examined as an indicator of indirect interaction.

The corpus of SEDE queries was also examined in relation to the question of what epistemic resources end-users produce through scripts on knowledge sharing platforms. As a first step, the highly structured nature of the SQL language that SEDE queries are written in was exploited to produce an overview. SQL *queries* retrieve data and operate on them. These can be made up of several different elements, but for the purposes of this study, the basic elements of interest are the commands that are known as *keywords*, the *identifiers* that are the names of objects in a database, and the *operators* that define how *keywords* and *identifiers* relate to each other. In the T-SQL standard adopted by SEDE, there are 179 *keywords*, the Stack Exchange database has 97 different *identifiers*, and there are 18 *operators* that may be used. With a table showing how much each keyword, identifier and operator is used providing an overview of SEDE, the queries were then examined through a process of frequency-inverse document frequency (TF-IDF) and k-means clustering to produce five clusters of 10 keywords and identifiers. This analysis was performed using the *SciKit Learn* library for Python. With clusters of queries identified, the associated queries were then ranked in terms of the number of stars assigned to them by other users. Finally, the most popular query for each cluster was examined as an epistemic resource in terms of the data chosen and operations performed, along with the title and description assigned by their authors.

Collective activity

In our analysis of the two chatrooms on the topic of SEDE scripting, we found that 1026 (38%) of the 2702 discussion threads in the corpus included more than one user. Graphing those threads to show the replies between users shows that the two chatrooms are only weakly connected, but that within each room, there are groupings and central users. Figure 1 shows a graph of reciprocal interactions between users in the corpus with the size of the nodes indicating the relative number of messages posted by a user and the thickness of the edges between nodes indicating the relative amount of reciprocal interaction.
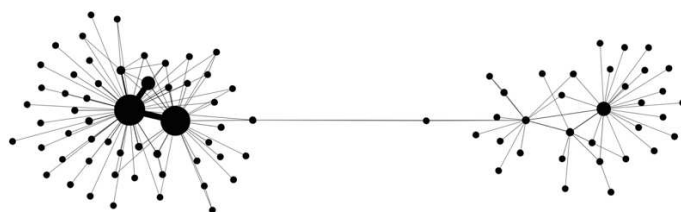

Figure 1: Network of reciprocal interactions between community members in SEDE chat threads

The presence of discussion threads along with groupings of users who have repeatedly replied to each other indicates the presence of direct interaction on the topic SEDE queries and suggests some collective activity. An examination of the content of recent threads shows a pattern of users reaching out for help to solve a non-functioning script or posting ideas for a script that they do not have the ability to write on their own. These posts are then often responded to by one or more of a small group of central users who are visible as larger nodes in Figure 1. They tend to provide suggestions for improving scripts or propose their own solutions to problems. While the average length of threads is short at 2.63 messages, some threads show more elaborate interactions where users share different solutions and post the results of testing scripting ideas. Since all scripts entered into SEDE are saved by the system, users also propose solutions based on other scripts and 'forking' or repurposing existing scripts makes up another form of collectivity in relation to the activity of scripting. Examining the corpus, we found that 1609 of 26,370 queries in the corpus had the same title as another query and 801 queries had the same description as another. This indicates that users do 'fork' the queries of other users to make use of them themselves. However, it is reasonable to assume that many if not most users who 'fork' queries and build on the work of others, change the name and description, meaning that this finding should be understood as an indicator that there is stigmergic interaction enabled by platform affordances rather than an assessment of the scale of that interaction. Taken together, these findings suggest that there are different types of collective activity associated with scripting ranging from attendance, when users are driven by individual goals and are not

engaged in joint work, to cooperation when 'forking' is accompanied by direct interaction in chats allowing users to establish and work towards shared goals.

Epistemic resources

Examining the corpus of queries themselves, we found that users had made use of 125 of the 179 available T-SQL *keywords* or commands with each used on average 2158 times in the corpus of 26,370 queries. Users also composed queries using all the 97 available *identifiers* for data contained in the Stack Exchange database with each used on average 357 times. Similarly, all 18 of the available *operators* were used at least once. This indicates that users produce a relatively broad selection of different queries that work with the range of data available from the platform in varied ways. However, some query elements were used more commonly than others (see Table 1).

Table 1: Most commonly used elements in Stack Overflow queries

T-SQL Keywords		Database Identifiers		Operators	
String	Occurrences	String	Occurrences	String	Occurences
as	36289	tags	2545	=	29444
and	18965	id	2359	--	11404
from	18778	reputation	2273	+	7575
select	17994	score	1830	>	3367
by	14481	creationDate	1684	*	2749
like	13864	name	1443	>=	2026
when	13575	postTypeId	1406	/	2013
then	13543	userId	1178	<	1725
where	12797	count	1126	-	1698
on	9691	ownerUserId	1120	<=	975

In Table 1, the ten most used of each type of query element are listed in terms of the number of times they appear in the corpus. As can be expected for a query-based information retrieval system (Wahl et al., 2017), the *keywords* and *operators* indicate a focus on selecting and comparing data. More revealing in terms of the epistemic resources that queries may become, however, is the choice of *identifiers*. These indicate an interest in the subject matter tags assigned to posts by users such as particular programming languages or techniques and on user-generated rankings including votes on posts and the reputation scores of users.

Examining the clusters produced through a frequency-inverse document frequency (TF-IDF) and k-means clustering analysis of the *keywords* and *identifiers* used in each query, the interest in subject matter and rankings is also visible (see Table 2).

Table 2: Most frequent clusters of terms used in queries

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
score	reputation	location	creationdate	text
id	users	reputation	id	comments
posts	id	like	posts	userid
tags	downvotes	lower	select	score
tagname	upvotes	users	count	comment
pt	count	upper	userid	link
join	select	desc	ph	like
posttags	displayname	row_number	postid	postid
select	votes	order	body	creationdate
owneruserid	user	user	votes	id

All five clusters include *identifiers* such as *score*, *reputation* and *votes* that indicate groupings of queries where rankings are made relevant. Cluster 1 includes the term *tags* indicating a group of queries that select data based on subject matter, but the other clusters include indicators of other interests. Cluster 2 indicates an interest in the upvotes and downvotes a user receives on their posts. Cluster 3 indicates an interest in the highest ranked users in a geographic location. Cluster 4 indicates an interest in the upvotes received by posts over time. Finally, cluster 5

indicates an interest in the comments made on posts in relation to their score. Each cluster indicates queries with different interests, but across clusters the terms suggest that different interests are all set in relation to user-generated rankings.

A common practice amongst writers of SEDE queries is that comments are imbedded in the body of queries. These comments are often preceded by two consecutive dashes (--), a symbol that instructs the T-SQL interpreter not to try to execute the text following it as if it were valid code. Examining the queries containing the terms found in each cluster that received the most 'stars' from users reveals comments that indicate the intended use of the script (see Table 3).

Table 3: Comment indicating intended use for most popular query in each cluster

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
"-- How many upvotes do I have for each tag?" (357 stars)	"-- How Unsung am I?" (242 stars)	"-- Top Users by Country" (74 stars)	"-- My Non Community Wiki Posts that earn the most Passive Reputation" (91 stars)	"-- StackOverflow Rank and Percentile" (127 stars)

As can be seen in Table 3, the comments found in the most popular query for each cluster confirm the different interests indicated by the terms in Table 2, but also underscore the dominant interest in scores and rankings. In general, the queries provide ways for users to understand their own scores and rankings in relation to different aspects such as topics, locations, experience on the platform, and badges awarded. For example, the query highlighted for cluster 3 displays top ranked users for specific geographic locations and the query for cluster 5 gives new users their rank when their reputation score is less than 200, the threshold at which the platform displays users in official ranking tables.

The query highlighted for cluster 1 allows the user to evaluate how many upvotes they need to receive before they are awarded a badge acknowledging their contribution to a particular knowledge domain (indicated with a tag). This interest in badges is also visible in name of the query highlighted for cluster 2 where 'Unsung hero' is the name of another badge. The query produces a list of a user's answers marked as accepted by question askers and the points that those answers contributed to the user's overall reputation score. Since question answerers invest time and effort in answering questions and may also aim to achieve a higher reputation ranking, they may be interested in examining which of the accepted answers they or other users have provided receive many upvotes and which do not.

The query highlighted for cluster 4 returns those posts for which a user receives the most 'passive reputation' - meaning posts where reputation score continues to be generated long after they are posted (for this query operationalised as 60 days). This gives an indication of those posts that have sustained value in the knowledge repository with new users finding them and indicating their usefulness by upvoting. This query indicates an interest in understanding the kinds of posts that are most likely to generate reputation score and thereby improve a user's ranking. However, revealing 'passive reputation' also offers users a way to understand the daily reputation increase cap for each post enforced by the platform. The query provides the basis for assessing the amount of reputation increase a user would have received had there been no cap on the amount of reputation score a post can generate each day.

Each of the queries in Table 3 produces an epistemic resource that is not available through the GUI features provided by the platform. However, rather than solely offering an overview of a particular aspect of the data, the queries highlighted for clusters 1, 2 and 4 provide resources for users to develop an understanding of how they might gain reputation points more effectively.

Discussion

In this paper, we have examined the phenomenon of end-users writing scripts that make use of the trace data on an online platform in different ways to those offered in standard Graphical User Interfaces (GUIs). Taking the programming knowledge sharing platform Stack Overflow as a context of investigation and the queries written for the Stack Exchange Data Explorer as the scripts in question, we have examined aspects of script writing along with scripts themselves. This work has focused on two aspects of the phenomenon asking how the writing and use of such scripts can be understood as collective activity and how they constitute epistemic resources.

With the aim of examining how SEDE query writing and use is a collective activity, we examined both the direct interactions of users as they discussed these practices in chatrooms and their indirect or stigmergic

interactions as they 'forked' each other's queries. Our analysis did not reveal the scope of either type of activity, but it did reveal the presence of it and suggest that collective activity takes place. Following the *Attendance*, *Coordination*, *Cooperation*, *Collaboration* (A3C) framework proposed by Jeong et al. (2017) and based on the findings of this study, we argue that the activity examined in this study constitutes collective activity that in some occasions can be characterized as *cooperative*. The presence of repeated reciprocal interactions between groups of users on the topic of query writing along with the 'forking' of queries themselves, suggests that "People have shared goals, but may also behave according to individual sub-goals", that "Processes are distributed, but can be shared on a limited basis", and that "Shared outcomes are salient, but they are more or less individual contributions" (Jeong et al., 2017, p.140). However, as with other instances of activity on large scale open platforms, it is likely that all four of the A3C forms of interaction take place in relation to SEDE with some users solely acting independently or *attending* as they 'fork' or write scripts without direct interaction with others. Meanwhile other users may take more collective responsibility and *collaborate* according to shared goals. Revealing the particulars of these different forms of interaction would require further research to examine the practices of users in detail, but the findings of this study suggest the presence of collective activity and evidence platform affordances allowing for such activity, thereby providing an orientation for that work.

Turning to the aim of better understanding what epistemic resources end-users produce through scripts on knowledge sharing platforms, the findings of this study reveal a particular interest in several aspects of the trace-data available in the Stack Exchange database. More specifically, the examined scripts demonstrate a dominant interest in points, rankings, and relative performance that cuts across interests such as topic, location, and particular groups of users. Following the argument made by Fischer et al. (2004) in relation to end-user development generally and Wahl et al. (2017) in relation to query based information retrieval, we argue that the aspects revealed are a source of insights about the knowledge practices on the platform. While previous research was primarily concerned with users' production of domain-specific knowledge (van Aalst, 2009; Wise & Schwarz, 2017), our study sheds light on complementary knowledge practices, as users draw on platform trace-data to "know their distributed communities and act within them" (Geiger & Ribes, 2011, p. 1). The results suggest that users produce epistemic resources that allow them and others to understand their performance on Stack Overflow in relation to others who have an interest in the same topic or live in the same area, but also provide ways to understand how one might most effectively contribute to the platform in terms of increasing one's reputation score. While the GUI of the platform provides an overview ranking of users based on reputation score, the SEDE queries indicate an interest amongst users to break down those rankings and produce detailed knowledge about their own and other's activity. Since collectively produced scripts are publicly available, they can also serve as resources for other users who observe scripting activity by pointing at metrics as a phenomenon of relevance for platform users. In this sense, such traces of scripting activity may contribute to the "alignment of individual goals and work processes to those of the community" (Jeong et al., 2017, p. 141). This argument is further supported by previous studies, which show that other mechanics such as received votes (Bornfeld & Rafaeli, 2019) and approaching a threshold for a new badge (Anderson et al., 2013) impact user behaviors on Stack Exchange platforms.

Taken together the findings of this study reveal the phenomenon of SEDE scripting on Stack Overflow to be a rich site of knowledge practices. Users interact directly on the topic of scripting queries though chatrooms, and indirectly, or in a stigmergic fashion, by 'forking' and building on each other's work. While Stack Overflow is a particular context with a user base that is perhaps more likely than in the case of other knowledge sharing platforms to have the necessary competencies to write and use scripts, the mechanics of voting and rankings are relatively common. Other Community Question Answering sites on a multitude of other topics run by Stack Exchange use the same mechanics and scripting interface. Other sites such as Quora and Reddit have similar points systems and have Application Programing Interfaces that make script writing possible. There are also tools available for writing browser scripts that modify the interface of nearly any website or platform. Given the presence of a scoring system and opportunities for script writing on a platform, it can be argued that similar use to that identified in this study is likely to be found. However, the contingencies of different platforms and the practices associated with them may lead to similar scripting activity that produces entirely different epistemic resources. Either way, as the findings of this study show, scripting provides insights into the details of knowledge practices on large online knowledge sharing platforms that may be obscured if only interaction with and through standard GUIs are considered.

References

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2013). Steering user behavior with badges. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*, 95–106. https://doi.org/10.1145/2488388.2488398

Atwood, J. (2010, June 13). *Introducing Stack Exchange Data Explorer*. Stack Overflow Blog. https://stackoverflow.blog/2010/06/13/introducing-stack-exchange-data-explorer/

Bornfeld, B., & Rafaeli, S. (2019). When interaction is valuable: Feedback, churn and survival on community question and answer Sites: The case of Stack Exchange. *Proceedings of the Annual Hawaii International Conference on System Sciences*, 789–799. https://doi.org/10.24251/HICSS.2019.096

Dillenbourg, P. (1999). What do you mean by collaborative learning? In P. Dillenbourg (Ed.), *Collaborative-learning: Cognitive and computational approaches* (pp. 1–19). Elsevier.

Fischer, G. (2016). Communities of interest: Learning through the interaction of multiple knowledge systems. *Proceedings of the 24th IRIS Conference*, *1*, 1–13.

Fischer, G., Giaccardi, E., Ye, Y., Sutcliffe, A. G., & Mehandjiev, N. (2004). Meta-design: A manifesto for end-user development. *Communications of the ACM*, *47*(9), 33–37. https://doi.org/10.1145/1015864.1015884

Fu, E. L. F., van Aalst, J., & Chan, C. K. K. (2016). Toward a classification of discourse patterns in asynchronous online discussions. *International Journal of Computer-Supported Collaborative Learning*, *11*(4), 441–478. https://doi.org/10.1007/s11412-016-9245-3

Geiger, R. S., & Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. *2011 44th Hawaii International Conference on System Sciences*, 1–10. https://doi.org/10.1109/HICSS.2011.455

Jeong, H., Cress, U., Moskaliuk, J., & Kimmerle, J. (2017). Joint interactions in large online knowledge communities: The A3C framework. *International Journal of Computer-Supported Collaborative Learning*, *12*(2), 133–151. https://doi.org/10.1007/s11412-017-9256-8

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation* (1st edition). Cambridge University Press.

Lindgren, S. (2020). *Data Theory: Interpretive Sociology and Computational Methods*. Wiley.

Nardi, B. A. (1993). *A small matter of programming: Perspectives on end user computing*. MIT Press.

Paavola, S., Lipponen, L., & Hakkarainen, K. (2004). Models of Innovative Knowledge Communities and Three Metaphors of Learning. *Review of Educational Research*, *74*(4), 557–576. https://doi.org/10.3102/00346543074004557

Stack Exchange. (n.d.-a). *Database schema documentation for the public data dump and SEDE*. Meta Stack Exchange. Retrieved June 15, 2021, from https://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede

Stack Exchange. (n.d.-b). *Stack Overflow*. Retrieved October 25, 2021, from https://stackoverflow.com

Stack Exchange. (n.d.-c). *What is reputation? How do I earn (and lose) it?* Stack Overflow. Retrieved March 30, 2021, from https://stackoverflow.com/help/whats-reputation

van Aalst, J. (2009). Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses. *International Journal of Computer-Supported Collaborative Learning*, *4*(3), 259–287. https://doi.org/10.1007/s11412-009-9069-5

Wahl, A. M., Endler, G., Schwab, P. K., Herbst, S., & Lenz, R. (2017). We can query more than we can tell: Facilitating collaboration through query-driven knowledge-sharing. *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 335–338. https://doi.org/10.1145/3022198.3026327

Wise, A. F., & Schwarz, B. B. (2017). Visions of CSCL: Eight provocations for the future of the field. *International Journal of Computer-Supported Collaborative Learning*, *12*(4), 423–467. https://doi.org/10.1007/s11412-017-9267-5