Exploring the Knowledge Curation Work of Wikidata

Timothy Kanke
School of Information
Florida State University
tjk11b@my.fsu.edu

ABSTRACT

The purpose of the proposed research is to explore how editors participate in Wikidata and how they organize their work. This research addresses the need to have greater knowledge of how Wikidata editors curate data for use. Three theories will be used as a methodological and conceptual framework for data collection and analysis: Activity Theory, Social Identity Theory, and Self-determination Theory. The study uses an exploratory case study methodology. An understanding of the activities in Wikidata, including the roles played, tools used, norms and rules followed, and solutions sought to address contradictions among different components of the activities will help inform communities wishing to contribute data to or reuse data from Wikidata. Furthermore, the findings of this study can go beyond understanding how Wikidata curates knowledge and potentially inform the design of other similar online production communities, scientific research institutional repositories, digital archives, and libraries.

Keywords

Online curation communities, Curation, Wikidata

ACM Reference format:

Timothy Kanke. 2018. Exploring the Knowledge Curation Work of Wikidata. In *Proceedings of Joint Conference on Digital Libraries (JCDL '18). ACM, New York, NY, USA, 14 pages.* https://doi.org/0

1. INTRODUCTION

This research explores how members participate in Wikidata and how they organize their work. At the time of conducting this study, there has not been an exploration of Wikidata members and their activities. This study addresses the need to have greater understanding of knowledge curation work in a large-scale peer-curated ontology. Wikidata is a free collaborative multilingual database collecting structured data for Wikimedia projects, that was launched in October 2012. It provides a centralized location for Wikimedia projects to query data allowing easier updating of information. According to Wikimedia September 2018 statistics, there are more than 50 million items in Wikidata that have been contributed by more than 50 thousand members with an active base of 19,133 contributors making at least 5 edits per month. A data quality survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO found that OpenCyc and Wikidata can be trusted the most on a knowledge base level [19]. The survey also found that Wikidata has the highest degree of schema completeness, population completeness, timeliness frequency. Finally, Wikidata is the most diverse in having labels in multiple languages. Although Wikidata is younger than other knowledge bases, the constant process of curating content in Wikidata has an advantage over the more periodically updated projects such as DBpedia [55]. Web-based knowledge bases are being used in diverse applications including Web search, natural language annotation, and translation. Another knowledge base, Freebase launched in 2007 to create structured data from Wikipedia and was later acquired by Google in 2010. The Knowledge Graph team at Google determined that Wikidata's fast growing and active community was "better-suited to lead an open collaborative knowledge base" [70 pp. 1420]. As part of the shutdown process a team moved a large portion of the Freebase data to Wikidata in 2015.

An understanding of the activities in Wikidata, including the roles played, tools used, norms and rules followed, and solutions sought to address contradictions among different components of the activities will help inform communities wishing to contribute data to or reuse data from Wikidata. Furthermore, the findings of this study can go beyond understanding how Wikidata curates knowledge and potentially inform the design of other similar online peer-production communities, scientific research institutional repositories, digital archives, and libraries. At this time, little research has been conducted in examining the knowledge curation work of Wikidata.

2. RESEARCH QUESTIONS

The study will address the following research questions:

1. What are the activities performed by members in Wikidata?
2. Who are the members participating in Wikidata?
3. Why do the members participate? What are the motivations of members contributing to Wikidata?
4. When do members' participate and how?
5. What is the division of labor within these activities? What are the roles played? What tools are being used

JCDL '18, June 3-7 2018, Fort Worth, TX, USA

https://doi.org/0

to perform these activities? What are the norms and rules regulating these activities?

6. How do the members negotiate and coordinate their work as a group and resolve contradictions that arise within and between activities?

3. BACKGROUND & RELATED WORK

3.1 *Knowledge Curation and Support Activities*

A few studies have looked at the knowledge curation activates performed in Wikidata. Some studies focus on the migration of data from other sources including drug-drug interactions [8] and Freebase [68]. The migration of data from Freebase to Wikidata developed an approach to perform migration and tools such as data preparation scripts and the Primary Sources Tool. The project created over 14 million new Wikidata statements helping increase the completeness and accuracy of Wikidata. Other studies compared Wikidata to Wikipedia. Steiner [63] found that Wikidata members use bots more than members in Wikipedia in a study of different language versions of Wikipedia and Wikidata. Wikidata has a lower occurrence of anonymous edits in comparison to Wikipedia. A majority of active members started on Wikipedia and have been gradually shifting their activity to Wikidata. Piscopo, et al. [53] compared the relationship between Wikidata and Wikipedia by analyzing external references. The study found that only a small number of direct reuse of sources occurs between the two Wikimedia projects. However, these references tend to point to the same domains. Finally, it was observed that Wikidata uses less Anglo-American-centric sources.

Piscopo, Phethean, and Simperl [52] conducted a study of how Wikidata member's forms of participation change overtime by interviewing eight Wikidata members who have had previous experience in Wikipedia. Using activity theory and the reader-to-leader framework, the study found that as a member's participation increased that the member's sense of responsibility, community interaction, number of tools used, and difficulty of tasks increased.

Most studies on Wikimedia knowledge curation and support activities focus on Wikipedia. There are many reasons for the sustained interest in studying Wikipedia including, long-term popularity, publicly available edit histories, rich structure, and identification of contributors [34]. Although Wikidata is not Wikipedia, both projects are created by the Wikimedia foundation and share members across the projects. Thus, the remainder of this literature review will discuss curation activities in Wikipedia. Kraut et al. [36] identified the challenges of online communities to become and remain successful. The challenges are starting new communities, dealing with newcomers, encouraging contribution, encouraging commitment, and regulating behavior. These five challenges provide a useful basis for a taxonomy of the social issues covered in the research literature.

Starting new communities. The process of starting a new community has "... a critical mass problem: the fledgling site doesn't yet have enough content to attract users and there are thus far too few users to create the content that might attract others" [36, p.3]. One of the challenges of studying the activities related to starting a new online community is that this requires a new community or the creation of an experimental community. Kraut et al. [36] note that few studies have been conducted in this challenging area, however, using theoretical arguments they identified six categories of design options: selection, sorting and filtering; community structure; content, tasks and activities; external communication; feedback and rewards; and presentation and framing. Although no study has yet to be identified within the literature focusing directly on Wikipedia and Wikidata, one study using the MediaWiki software focuses on the content, tasks, and activities related to the effectiveness of seeding content in a wiki [62]. The results of the experiment showed that groups with seeded content added more structural information than the blank group. However, participants in the blank group provided more facts. This motivation can be explained by self-determination theory where autonomy is an important factor of intrinsic motivation. In regards to whether seeded content influence, users do look at the previous material to guide their own behavior.

Dealing with newcomers. Dealing with newcomers is an important issue for online communities because newcomers are a source of innovation of new ideas, work procedures, and other resources, as well as, compensating for the inevitable turnover of current members. Some of the topics covered in the literature are an awareness of the project, ways to solicit contributions, editing activities, and providing feedback.

Many issues arise due to the fact newcomers do not have knowledge or experience with an existing community's culture and activities. Antin and Cheshire [3] argued that characterizing all readers as "free-riders" is unfair because readers may not have complete information about participation options. Furthermore, reading is a form of contribution that is a gateway to active participation. 165 students from a large West Coast public university participated in the survey. The authors found that few participants knew about specific participation policies, even though they knew that one could contribute to Wikipedia. This incomplete operational knowledge is used as support that readers are not free-riders. One way to increase awareness is to launch special projects to invite participation. In 2010, the Wikimedia Foundation launched the Wikipedia Public Policy Initiative that increased contributions to policy-related articles by conducting compulsory university course assignments. Lampe, Obar, Ozkaya, Zube, and Velasquez [42] surveyed 185 students on their impressions of the activity and any intentions for future contributions. Two variables that may predict the likelihood of future contributions are the level of reported class engagement and global visibility awareness. Of the 615 students that participated in the initiative, only 25 people have edited Wikipedia after the class ended. Although this is a low return rate it is at a better rate than the usual rate of normal contributors of 0.0002%.

Once there is awareness of the online community the discussion turns to how does one invite newcomers to

contribute. Choi, Alexander, Kraut, and Levine [13] conducted a study about newcomer retention through socialization tactics used in WikiProjects. Seven socialization tactics are identified: invitations to join, welcome messages, requests to work on project-related tasks, offers of assistance, positive feedback on new participant's work, constructive criticism, and personal-related comments. The authors noted that this study is consistent with prior research in that the number of responses the newcomer receives is associated with the number of edits regardless of the nature of the response. In particular, constructive criticism is the most effective in retaining newcomers. However, this study differs from the literature with personalized tactics appearing to be more effective than standardized tactics. Standardized tactics are structured learning experiences that are designed to encourage newcomers in a consistent manner. Personalized tactics tailor the experience to the individual.

Some research has been conducted on the editing activities of newcomers. For example, Antin, Cheshire, and Nov [4] examined technology-mediated social participation (TMSP) of new contributors within their first 6 months of activity. The definition of contributors is different than many other studies because the emphasis is on moderate sustained activity (i.e. 10 revisions per 30-day period) among new members instead of focusing on highly active contributors over extended periods of time. The authors reviewed TMSP theories to inform their study. They found that most contributors experience a "honeymoon effect" and decrease the amount of activity over time, the earlier a contributor diversifies the type of activity tends to perform more advanced activities later, and that contributors who perform administrative tasks early will tend to perform more administrative tasks later.

The method of providing feedback can have an influence on a newcomer. Ciampaglia and Taraborelli [14] conducted an experiment using MoodBar, a lightweight socialization tool, to encourage activity and retention of newcomers for Wikipedia. MoodBar elicits feedback from newcomers about their editing experiences and is displayed on a dashboard as a mood (i.e. happy, sad, or confused). MoodBar is used to report early editing experiences with median times of happy and confused at 30 minutes and sad at about 2 hours. Individuals that received feedback where about 9 times more productive than people who never sent feedback. They found that early feedback encouraged higher short-term productivity and improved long-term retention of newcomers. However, if the feedback is provided in a negative manner it can deter newcomers. Suh, Convertino, Chi, and Pirolli [65] found an increase of reverts-per-edit, the number of protected pages and resistance from "deletionist" editors have slowed down the growth of Wikipedia. Furthermore, Halfaker, Kittur, and Riedl [24] found that edit reverts can be demotivating to newcomers especially if the revert is performed by a more experienced editor. Although this process chases away newcomers, the people who stay create work of greater quality but in less quantity. This has been found to be true for the creation of policies as well. Halfaker, Geiger, Morgan, and Riedl [23] identified two consequences of "policy calcification" that influence newcomer socialization and retention. First, newcomers experience policy edits rejected at a higher rate than what has occurred in the past. Second, newcomers are contributing more to essays that are not as likely to be reverted. However, essays contain informal norms that are superseded by formal norms. The researchers also noted that the increased use of automated revert tools that only accept or reject has had an unintended result in discouraging newcomers to contribute.

Encouraging contribution. Encouraging contribution is an important factor for an online community's success because a group's existence is built on the activity of members contributing resources. The literature in this category can be divided into three topics: selection, sorting, highlighting of tasks; intrinsic and extrinsic rewards as motivators; and community structure.

Selection, sorting, highlighting of tasks. Many articles cover the selection, sorting, and highlighting of tasks in Wikipedia focus on how Wikiproject groups provide a structure for these tasks. Tinati, Luczak-Roesch, Shadbolt, and Hall [69] examined WikiProjects as a perspective for understanding activities occurring in Wikipedia. The data set is the 3.2 million unique wiki pages (i.e. articles and talk pages) that are connected to 618 active WikiProjects. In the comparison of projects, the study found a normal distribution of articles, Talk pages, and a number of members. There is a correlation between the number of Talk entries on the project page and the number of articles edited. The authors propose that WikiProjects are a suitable proxy for studying contribution and social activity in Wikipedia.

Kittur, Pendleton, and Kraut [35] examined the influence of group structure provided by WikiProjects on Wikipedia members. A comparison of members that join a project was compared to non-project members within a set of 73 WikiProjects. The authors found that after joining a project member tend to focus attention towards group related tasks, as well as, their behaviors will most likely shift their contributions from production work to coordination activities. Also, members that join a group tend to have an increase in good citizenship and maintenance behaviors. However, project membership has negligible influence on the amount of editing activity.

Gilbert, Morgan, McDonald, and Zachry [22] studied the complete revision history of every article linked to a Hot Articles page for the time frame of 90 days prior and 90 days after activation of the Hot Articles bot. The sample included a combined total of 9,961 articles from 7 WikiProjects. The results indicate that links posted in the Hot Articles list received a greater number of edits and unique editors on average than links posted to talk pages. Also, there may be a distinction between the types of edits for members versus non-members, however, the introduction of the Hot Articles bot links did not produce a significant difference. The authors stated that a factor of this result may be due to how membership was measured in this study.

Morgan, Gilbert, McDonald, and Zachry [49] examined how participation in WikiProject groups have changed over the life of Wikipedia. The WikiProjects that are of concern are the ones that fall outside of the normal subject domain focused groups.

Since 2007, the number of WikiProjects has been in decline. Although the number of alternative WikiProjects has been increasing, conventional projects are still the most active projects. The authors used 6 categories for alternative WikiProjects: editing work, meta-content work, social and community support actions, collaborative actions and disposition, border patrol, and administrative actions. WikiProjects have a "pivotal role" in coordinating completion of tasks. A future study could examine if this is true in other wiki-type projects thus suggesting a general trend within open collaborations. The authors suggest 4 design elements for supporting WikiProjects and other similar groups: general requirements, socially intelligent task routing, socially intelligent task routing, social translucence visualizations, and leaderboards.

Zhu, Kraut, & Kittur and [76] studied how group identification and direction setting influence people to complete group tasks. A total of 26 WikiProjects that use Wikipedia's Collaborations of the Week (COTW) with 5 or more editors and distinct topics were used for analysis. The authors proposed 3 hypotheses and created separate tests for each. The first test examines the direct effects of goal setting and found that group goals publicized via COTW have a strong influence on group members in comparison to non-self-identified editors that contribute to a task under a group's domain. The second test examines the spillover effect of goal setting and found that the effects of goals influence non-goal related tasks. The third test examines the effects of social modeling and found that group role models in COTW influence editors to exhibit similar behaviors.

Morgan, Gilbert, Zachry, and McDonald [48] investigated the role of explicit coordination present in WikiProject talk pages. The authors note that threaded email and forum discussions allow "loosely-affiliated individuals" the ability to work asynchronously on interdependent tasks. However, they identified that "the role of variables such as medium, task type, task complexity and group size on explicit coordination" have had little research. Similar to other researchers, level of talk page activity is considered a strong indicator of article quality. The data set is a random sample of 138 highly active WikiProjects between July 2011 and July 2012. The preliminary findings show that 73% of posts contain explicit request or proposal for coordination but few posts explicitly requested collaborative editing.

Ung and Dalle [70] investigated the relationship between project-based coordination activity and collective production behaviors. The authors sampled French Wikipedia for WikiProjects and removed pages that redirect to other pages, projects that were not linked to any articles, and pages with limited activity. A time series of weekly coordinated activity was measured by counting the number of edits on the project-pages per week. 98% of projects have at least one burst of activity often with one of the bursts near the beginning of the life of a project. 66% of the projects show two or more activity bursts with an average coordination activity representing 69% of the activity. The results indicated that coordination activity in project pages is reflected in the article's editing activity. Their analysis also showed that managerial and coordinated activity of leaders influences the contributor activity.

Morgan, Gilbert, McDonald, and Zachry [48] examined a one-year sample of 788 talk pages from 138 WikiProjects with the purpose of understanding how editors, both self-identified group members, and non-group members, use group workspaces to propose, prioritize, and perform work. Non-member participation in WikiProjects was much higher than expected with 54% of the sampled posts being initiated by non-members. Also, non-members posted 37% of all messages in the sample. The authors found that non-members post certain types of requests (i.e. invitations to join other projects, requests for others to perform edits for them, and requests for external discussion participation), however, the difference between members and non-members posting behaviors lack overall significance. Members tend to reply more to fellow members than to non-members possibly due to shared history or common bond. However, members are not more likely to perform work proposed by other members over non-member requests.

Forte et al. (2012) analyzed editing activity to investigate how nested organization structures support collaboration. A typology of group function informed the analysis of the interviews by focusing on three group functions (i.e. production, group well-being, and member support) and four modes of production (i.e. inception, problem-solving, conflict resolution, execution). Interviews were conducted in two stages with many people belonging to WikiProject Military History. A quantitative analysis looked at the edit histories of 379 projects from 2001 to 2008. The activity of each project is measured by the average number of edits to a project page per month. This graph of the data does not have the expected power law distribution with a long tail. Instead, there are many projects with a moderate amount of activity and few projects with little activity. The authors found that WikiProjects help coordinate tasks and produce articles, as well as, support community members by provides a place for collaboration, socialize and network, protect editor's work, and structuring contribution opportunities. In contrast, a study that did not use Wikiprojects as a population had different findings. Krieger, Stark, and Klemmer [37] conducted a 3-month qualitative online ethnographic study examining how Wikipedia editors organize their actions as well as other editor's actions. The authors identified 5 principles insights about task management for Wikipedia: bottom-up structure, template tags do not function well as next actions, lack of triage, disconnect between individual and site tasks, and lack of support for contextual discovery.

Some tasks are started due to a current event such as the death of a notable person. Keegan and Brubaker [29] investigated death work on Wikipedia. The "Notability" policy restricts the nature of the content of a biographical article when a person is living. However, this policy is no longer valid once the person is deceased. Wikipedians use journalistic obituary writing skills to document social memory and history. One difference is that many Wikipedia article exist before the death of a person, thus go through a transformation. The authors compared the number of pre- and post-mortem revisions and

found that people who are popular in life are popular in death. Collaboration networks change upon the death date including editors focusing on a few articles instead of contributing across many articles. An expect spike of activity occurs on the death day and decays rapidly until returning pre-mortem ranges within a month.

Intrinsic and extrinsic rewards as motivators. Intrinsic and extrinsic rewards have been studied as potential motivators for contribution. Nov [50] found that Wikipedians' motivations to contribute are intrinsic such as ideological commitment for the project and for the fun of it. Other motivators such as social reasons, career advancement, and protection were not strong indicators. Oreg and Nov [51] found that Wikipedia editors are motivated more by altruistic motives more than reputation-gaining. Also, relationships between personal values, including achievement, self-direction, benevolence, and universalism, and their motivations to contribute. Shao [61] developed an analytical framework of user-generated media that identified three reasons to participate: "consume content for fulfilling their information, entertainment, and mood management needs" (p.7); enhancing social connections and producing contents for self-expression. Cho et al. (2010) examined motivations, internal cognitive beliefs, social-relational factors, and knowledge-sharing intentions of Wikipedia editors. The authors found that attitudes, knowledge self-efficacy, and a basic norm of reciprocity have a significant relationship with knowledge-sharing intentions. Altruism (an intrinsic motivator) was found to be positively related to attitudes toward knowledge sharing, while reputation (an extrinsic motivator) is not. Finally, they found that a sense of belonging is related to knowledge-sharing intentions indirectly through many motivational and social factors including altruism, norms, knowledge self-efficacy, and reciprocity. Intrinsic motivators also play a role in unwanted contribution to Wikipedia. Shachaf and Hara [60] studied Hebrew Wikipedia trolls' behaviors and motivations and found that boredom, attention seeking, and revenge to motivate trolls. They exhibit an intrinsic motivator of finding pleasure from causing damage performed with anonymous or hidden identities.

Although intrinsic motivators are found to be the primary motivation for contribution, there is at least one extrinsic reward that is effective. Kriplean, Beschastnikh, and McDonald [38] analyzed English Wikipedia to identify barnstars that were awarded to create a classification of how they are used. Barnstars are acknowledgment-based rewards that usually acknowledge some form of work as being exemplary. They also serve to lessen social strife, encourage new members, and foster competition. A barnstar can belong to more than one category. The categories are editing work, social and community support actions, border patrol, administrative, collaborative actions and dispositions, and meta-content work.

Community structure. Forte and Bruckman [18] noted that Wikipedia's minimal user privileges and ability to provide desirable activities help encourage contribution. However, this is not to say that there is not a structure to Wikipedia's community. Arazy, Ortega, Nov, Yeo, and Balila [7] identified six organizational levels ranging from unregistered non-community members to "benevolent dictator" Jimmy Wales. The structure has evolved over time to include technical administration, border patrol, quality insurance, and QA technicians. These positions were necessary to handle the increase of contributions and ensure the sustainability of the project.

Ransbotham and Kane [55] investigated how membership turnover affects collaborative projects. The authors discuss several views on membership turnover from the least amount of people leaving an organization is best to harvest knowledge quickly and have a quick turnover. A more moderate approach is the focus of the hypothesis that a membership turnover relates in a curvilinear way to the performance of both knowledge creation and knowledge retention to a point and then impairing it thereafter. The sample is articles from 2001 to 2008 that had attained featured article status. The individual revisions, minus edits performed by bots, are aggregated to monthly observations of editing activity. The control variables are article length, section depth, external references, internal references, reading complexity score, and multimedia intensity. Four logit statistical models are tested, and the results show the expected curvilinear shape. Interestingly the authors refer to Wikipedia as being social media that is contrary to the classification by other articles.

Robert and Romero [56] analyzed a data set of 4,317 articles connected to WikiProject Film to understand the effects of crowd size and diversity on crowd performance. The measures are article quality, topical diversity, inner workload diversity, outer workload, outer workload diversity, and crowd size. The study found that crowd size that is diverse leads to better performance. However, if the diverse group is small, then a similar less diverse group will outperform the diverse group. The design implications of this study suggest that recommendation systems should consider the number of members and current diversity of members in regard to how new members would change the ratio in addition to the current practice of assessing the individual characteristics of expertise and experience. Arazy and Nov [5] found that "global inequality exerts a significant positive impact on article quality, while the effect of local inequality is indirect and is mediated by coordination." The authors propose 3 practical concepts: A team should have an unequal composition, that is a combination of highly active and relatively inactive people; entry barriers should be low, and mechanisms should be in place for coordinating activities with uneven team composition.

However, not all inequalities have a positive impact on Wikipedia. Lam et al. [41] identified a large gender gap among members and a corresponding gender-oriented content disparity in articles. They found evidence hinting at a female participation resistant culture including a higher percentage of edit reverts during the early part of tenure and differences in social engagement. Collier and Bear [15] conducted a survey of female Wikipedia users and unearthed similar findings. Female contributors are less likely to contribute because they prefer to share and collaborate rather than delete and change other's work. The high level of conflict involved in the editing, debating, and defending process of articles were found to be a detractor of

contribution. Finally, they are less likely to contribute due to lower confidence in the perceived value of their contribution and their overall perceived expertise.

Encouraging commitment. Encouraging commitment is important in retention of membership because the greater a person's feelings of attachment towards a group the more likely a member will stay an active contributor of the community. This concept is closely related to encouraging contribution and dealing with newcomers. To differentiate this category, we will discuss literature that has a focus on the path, or lack thereof, to expert Wikipedian.

Balestra, Cheshire, Arazy, and Nov [9] conducted a two-part survey investigating the motivational paths of newcomer Wikipedia editors. The authors note that the literature states that motivations either increase when contributors move from the periphery to the core or decrease overall over time. This study found different motivations change in different ways. That is non-instrumental motives (collective, self-expression, and fun) decreased significantly over time. This suggests that newcomer's experiences are not altogether positive. Meanwhile, instrumental motives did not change over time (e.g. social motives increasing marginally). This suggests that social aspects of Wikipedia gain importance over time. As far as the trajectory of the career path, Arazy et al. [7] found that there is neither formally defined career paths nor a linear progression through the roles in Wikipedia. Contrary to prior literature, it was noted that many participants moved directly from entry levels to core roles and that some members showed a decrease in responsibilities from core activities to the periphery.

Wikipedia's diversity in tenure appears to have positive effects. Chen, Ren, and Riedl [12] studied the effects of group experience diversity on the amount of work accomplished and group membership numbers. Topical WikiProjects focus is to improve articles within a topic area by expanding content, unifying writing style, quality peer review. The project member lists were used to identify participating editors. The independent variables are Wikipedia tenure (oppose to specific project tenure) and topic interests. Topic interests were identified by the pages that editors contributed work excluding "incidental" edits (e.g. spelling correction). The dependent variables are the amount of work accomplished (cognitive performance), and member withdrawal from the group (group affective performance). The control variables are quarter index (length of time a project in quarter years), project size, project scope, and level of controversy (measured by number of reverts). The study's findings noted that increased diversity in experience with Wikipedia increases group productivity and decreases member withdrawal. However, if a group has either extremely low or high tenure, then the number of members that withdrawal from the group doubles. The authors note that interest diversification shows only positive effects.

Although Wikipedia is known for being egalitarian with providing an opportunity for anyone to contribute, the "In the News" stories are structured in a different manner. Anyone can nominate a story and there is an open discussion. However, only editors can vote on which stories will be chosen. Keegan, Gergle, and Contractor [31] divided the editors into 3 groups that were sampled in a 3-month time-span: elite editors made more than 10 contributions, middle editors made between 2-10 contributions, and drive-by editors made only one contribution. Elite editors support is influential but not any more influential than a middle editor on the promotion of articles. Elite editors and drive-by editors have a similar influence on supporting contested articles. A difference occurs with preventing nominations from being promoted. Elite editors have more influence than middle and drive-by editors. Thus, "In the News" stories exhibit "one-sided gatekeeping" were elite editors can block inappropriate stories, but do not have any more influence than other editors in terms of promoting stories.

Finally, the amount of diversity in activity can be an indicator of an editor's longevity. Wang, Chen, Ren, and Riedl [72] investigated what type of behaviors are related to members decreasing contributions to a group and the implications of their withdrawal behavior. Two "trade-offs" are investigated: increasing productivity and withdrawal; subgroups and the larger community. Productivity is measured by number of edits. Withdrawal is measured by the evidence of active editors ceasing contribution. The factors that affect productivity and withdrawal are tenure, tenure dissimilarity, past productivity, concurrent projects, and communication / social integration. The results show a negative relationship between tenure and both productivity and withdrawal. People involved in multiple projects contributed less, however, they were less likely to withdraw from any project. People who engaged in communication with others (inside and outside of a given project) contribute more than less social editors.

Regulating behavior. Regulating behavior is an important challenge for many communities. Online communities may have more challenges than traditional communities due to anonymity, ease of entry/exit, and textual communication. In this section regulating behavior is discussed in the context of coordinating activities, editing conflicts, leadership influences, and cultural influences.

Coordinating activities. Kittur and Kraut [34] compared the findings of the development and effectiveness of coordination mechanisms in Wikipedia to other wikis hosted on Wikia. Wikia has 6811 publicly available wikis that use the same software platform as Wikipedia thus removing differences in the structure of the platform. The top ten wikis account for 35% of all revisions made to all wikis. The lifespan of projects distribution is skewed with few communities becoming highly successful. These two factors are consistent with other online production groups. The wikis are like Wikipedia in that the amount of communication increases over time and stabilize. The group structure is also similar with much of the work accomplished by a core group. In regard to Wikipedia, Forte, Larco, and Bruckman [20] noted that the community structure was originally less formal during the early years. As the community grew governance became increasingly more decentralized over time due to the fact it had become difficult to achieve consensus.

Keegan et al. [31] focused on the interaction between two approaches of studying Wikipedia's collaboration processes: editor-focused attributes and article-focused features. The authors chose articles about commercial airline disasters since this type occurs frequently enough to produce a suitable sample (n=249) and these articles are newsworthy enough to involve the initial high amount of activity when the articles are created. A statistical model was used to compare the editor-focused attributes and article-focused features. They found certain article's coordination demands can influence an editor to seek or avoid these articles depending on the type of editor. Experienced editors tend to concentrate their contribution to fewer articles, as well as, are attracted (or repelled) from certain types of articles.

Jesus, Schwartz, and Lehmann [25] found that within the categories of philosophy and physics editors will form cliques of at least three editors and will share editing on at least five articles. Their analysis could identify controversies on topics such as intelligent design, the validity of intelligence tests, and global warming and that controversy are not necessarily confined to one article. The use of cliques appears to have an influence on regulating editing behaviors. Article quality often becomes a goal of these groups of editors to achieve featured article status.

Group characteristics. Arazy, Nov, Patterson, and Yeo [6] identified and examined how content and administrative focused groups influence quality. Content focused groups tend to center around a few select topics and have lower levels of commitment and identification with the greater Wikipedia community. This group produces higher quality articles in comparison to administrative focused groups. Administrative focused groups have a higher participation, commitment, and identity that is dispersed around various topics. This group enhances quality through reducing task conflict. This reduction of varied opinions on what needs to be done reduces the negative effects on quality. However, the elimination of task conflict would remove diversity. A lack of diversity can have detrimental effects on article quality.

The style of coordination is also important. Kittur and Kraut [33] examined the coordination of a group of editors has an influence on quality. Two types of coordination were identified. Explicit coordination is when the editors plan the article through communication. Implicit coordination is when a subset of editors structures the article in the initial stages. The application of appropriate coordination techniques determines if the addition of editors is helpful or harmful to article quality. In the beginning stages of article development, both types of coordination improved quality. Over an article's development, implicit coordination was more helpful when many editors contributed by concentrating the work.

Anthony, Smith, and Williamson [2] findings show that contributions with high reliability are from registered editors, who are motivated by reputation and commitment, as well as, anonymous single-time contributors. This mixture of contribution levels having an influence on quality also been observed by Arazy and Nov [5]. They found that Wikipedia-wide inequality of contribution, due to differences in commitment, interest, and involvement, exerts a significant positive impact on article quality. However, the effect of article-specific inequality of contribution is indirect and is mediated by coordination. Ransbotham and Kane [55] found that a moderate level of turnover, as well as a mix of new and long-term editors, are beneficial to article quality. This finding diverges from the long-held belief that long-term retention is essential for online collaboration.

Duguid [16] took two concepts from open source software peer-production and applied them to the work done in Wikipedia. The two concepts are that the higher levels of membership participation lead to higher quality and over time good-quality content will remain while bad-quality content will be removed. However, he determined that the highly democratic nature of Wikipedia makes it difficult to apply these two open source software production concepts. One study found that the group size has a positive effect on the effectiveness of the work being produced [11]. However, Warncke-Wang, Ayukaev, Hecht, and Terveen [73] found that an increasing the number of people per article is associated with a slower increase in quality.

Editing processes. Jones [26] states that the revision process in Wikipedia is unique in comparison to the revision patterns of academic and business writing. He examined the revision histories of 10 articles with half of them having featured status. All the articles showed a higher percentage of new material additions compared to deletions and text-based rearrangement revisions. The non-featured articles had fewer surface revisions and were dominated by content revisions. Projects with a clear purpose are connected to higher levels of quality. Finally, Projects that are well structured are more likely to succeed [73]. However, Adamic et al. [1] findings show a small but significant positive correlation between focus and quality when it comes to individuals. Individuals who limit their focus tend to produce higher quality work, but a very narrow focus are on average less well recognized among peers.

The use of features in Wikipedia also differentiate it from other styles of writing. Two studies, Stvilia, Twindale, Smith, and Gasser [64] and Ehmann, Large, and Beheshti [17] analyzed the discussions of quality on Talk pages and the quality of the articles. The Talk pages are an important resource to understand how the editors define quality and the process of improving the quality of articles. The results showed a strong connection between the discussions of quality and the quality of the articles and that issues of quality are important to the Wikipedia community.

Finally, reverts have an influence on the contributions of editors [24]. Reverts are important to maintaining quality by fixing mistakes, repairing vandalism, and helping enforce the policy. Halfaker et al. [24] measured motivation indirectly with effects that may be caused by changes in motivation, including reducing the amount of or ceasing contribution. Newcomers that are reverted by an editor with more experience in Wikipedia are more likely to withdraw or decrease in the quantity of work. However, a newcomer who stays used this learning experience to increase the quality and productivity of contributions.

Editing conflicts. Kittur and Kraut [34] found that the increase of contributors is associated with an increase in conflict. However, if the size of the wiki increases the number of conflicts decreases. This may be due to editors' perceptions of editors are formed in early interactions can have a lasting effect on how future work is perceived. Marlow and Dabbish [44] found that good first impressions have a lasting effect on how future work is perceived. While negative first impressions appear to make participants more sensitive to correcting bad work but did not have enduring perceptions.

Autonomous editing programs (bots) and assisted editing tools have changed the nature of editing and administration of Wikipedia. Geiger and Ribes [21] state the use of bots, scripts, and other tools began in late 2006. The first notable bot, RamBot, imported city and town census data into articles. By 2009, over 12 percent of the edits were performed by assisted tools. Another notable bot, Huggle, is used to fight vandalism. This piece of software can rank edits and suggest the next edit a person will want to check for vandalism. The rank is decided by several factors including significant removal of text, users whose edits have been previously reverted, and pages replaced with blank text. The software can also be used to predict outcomes of behaviors. Although bots have many beneficial uses, Vrandecic [71] states that a side effect of bots is that they are prominent in editing Wikipedia articles thus decrease editor engagement. It is hoped that Wikidata will reduce the amount of Wikipedia bot activity and result in higher visibility of human editing activity.

Leadership influences. Leadership has some influence on regulating behaviors in Wikipedia especially when the behaviors are task specific. Keegan and Gergle [30] studied the regulating effects of editor seniority. Elite editors support is influential but not any more influential than a middle editor on the promotion of articles. Elite editors and drive-by editors have a similar influence on supporting contested articles. A difference occurs with preventing nominations from being promoted. Elite editors have more influence than middle and drive-by editors. Thus, "In the News" stories exhibit one-sided gatekeeping were elite editors can block inappropriate stories, but do not have any more influence than other editors in terms of promoting stories.

Zhu, Kraut, and Kittur [75] used path-goal theory to explain the process by which leaders influence others' behaviors. The authors propose that task-based leadership, opposed to traditional vertical leadership, may be more useful in describing leadership roles of online communities. The findings are that leadership performed at all levels within a Wikiproject significantly influenced other members' editing behaviors. Also, positive person-focused leaders were more effective in motivating others, whereas aversive leaders decreased the number of contributions. Finally, people who were identified as leaders were, in general, more influential than regular members. In a related project, Zhu et al. [76] found that group goals publicized via Wikipedia's Collaborations of the Week have a strong influence on group members in comparison to non-self-identified editors that contribute to a task under a group's domain. It was found that goals have a spillover effect and influence other non-goal related tasks. It was also found that group role models in Wikipedia's Collaborations of the Week influence editors to exhibit similar behaviors including anti-vandalism editing activities.

Cultural influences. Pfeil, Zaphiris, and Ang [54] tested an assumption that there is a correlation between language and culture by choosing some of the Wikis that are the primary language for many cultures. The topic of game was analyzed from French, German, Japanese, and Dutch sites. The cultural dimensions are power distance, collectivism versus individualism, femininity versus masculinity, and uncertainty avoidance. Power distance is "the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally." The results indicated that there is a correlation between the power distance index and the number of deletion acts. Collectivism versus individualism refers to the extent that members of a culture look after a larger portion of their society or limit the care to themselves and their immediate family. A correlation between individualism index and number of corrective actions. Femininity versus masculinity describes the gender roles at the cultural and individual levels. A higher masculinity index score correlates to a decrease in the number of contributions in corrective categories.

3.3 Theoretical Frameworks

To conceptualize the organization of and activities around Wikidata, the study used Activity Theory, Social Identity Theory, and Self-determination Theory as a methodological and conceptual framework for data collection and analysis (see Fig. 1).

Activity Theory. Activity theory provides a hierarchical structure for studying and analyzing the activities of different communities with activity being the basic unit of analysis. The activity system model focuses on collective group activities that are performed by individual subject contributions. There are three elements: subject, object, and community. Each of these element's interactions is mediated by a special type of means. The subject-object interaction is mediated by tools. Subject-community interaction is mediated by rules. Community-object interactions are mediated by the division of labor [27]. One key tenet of the framework is that activity systems undergo a contradictory development process. These contradictions can be divided into four categories: inner contradictions of the activity system's components, contradictions between components, contradictions in the relationship between the forms of the system and its object/outcome, and contradictions within a network of activity systems [27]. There are six principles of activity theory: unity of consciousness and activity, object-orientedness, hierarchical structure of activity, internalization-externalization, mediation, and development. The concepts of internalization and externalization provide a lens to understanding a community's routinization and formulation of new tools, rules, and mechanisms for ontology development and maintenance. Activity theory enables connecting individual and community ontology development and maintenance activities to

the activities of Wikidata and the data curation processes in a social context.

Social Identity Theory. Social identity theory is used to investigate the direct effects of goal setting, spillover effect of goal setting, and effects of social modeling. Social identity theory was developed in the 1970s by Tajfel and his colleagues [76]. Social identity is that a person's self-esteem is derived from perceived membership in a group. Social identity theory focuses on intergroup behaviors by examining the perceived group status, legitimacy, and stability differences, as well as the perceived ability for someone to change group membership [66, 67]. In Zhu, Kraut, and Kittur [76] study of WikiProjects in Wikipedia, they noted that several behaviors are associated with a social identity that is beneficial to the group including cooperation, effort, beneficial decision making, intrinsic motivation, task performance, and information sharing.

Self-determination Theory. Self-determination theory is used to investigate the degree of self-motivation and self-determination of an individual. There are two types of motivations behind an individual choosing to perform an activity: extrinsic and intrinsic [57]. Extrinsic motivators are external rewards such as fame and monetary gain, as well as, the desire to avoid punishment. Intrinsic motivators are internal rewards that are performed due to a self-determined interest or is found to be enjoyable. Three basic psychological needs motivate an individual to engage in activities. Competence is the desire of an individual to seek mastery of an experience and control the outcome. Relatedness is the need to interact with others. Finally, autonomy is the need to be the causal agent of one's own life.

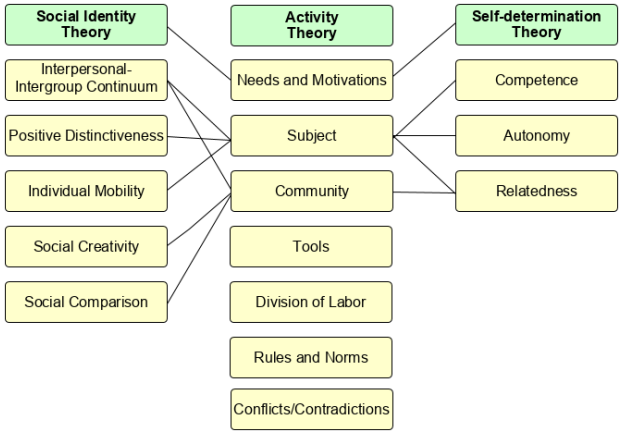

Figure 1. Theory Map

4. METHODOLOGY

The study will use an exploratory case study methodology. Yin [74] defines case study as an in-depth investigation of a phenomenon within its context where the boundaries of the phenomenon and context may not be clear. He elaborates on the characteristics of case studies that they are suited for coping with situations that have more variables of interest than data points, rely on multiple sources of data, and benefits from the theoretical framework in the design, data collection, and analysis.

The study has an embedded single-case design. Two sources of data are collected and analyzed. In particular, the study will use a qualitative content analysis to analyze documentation from main Wikidata talk and help pages, WikiProject pages, user pages, and property talk pages. The second source of data will be collected through interviews of Wikidata participants. To conceptualize the organization of and activities around Wikidata and interpret findings, the study will use activity theory, social identity theory, and self-determination theory as a methodological and conceptual framework for data collection and analysis.

The process of case studies can be grouped into five stages: plan, design, prepare, collect, analyze, and share [74]. These stages are performed in a linear iterative fashion. This study has an embedded single-case design. A single-case design focuses on one phenomenon and its context. An embedded design has the potential to enhance the insights of a single case by providing more data points and maintain focus on the goals of the study. Two sources of data are collected and analyzed. Documentation from main Wikidata talk and help pages, WikiProject pages, user pages, and property talk pages will be used for qualitative content analysis. The second source of data will be collected through interviews of Wikidata participants.

This proposed study is an embedded single-case design with two units of analysis: talk page threads and individual Wikidata participants. Activity theory is the theoretical framework used in the design, data collection and data analysis. Data and method triangulation is used to gain a better understanding of the Wikidata activities. The sources of the data and chosen methods are closely linked: documents in the form of project discussion threads are investigated with content analysis and the views of individual participants are collected via interviews.

Qualitative Content Analysis. The qualitative content analysis method that describes the meaning of data by reducing it to categories of a coding frame in a systematic yet flexible manner [45, 56, 59]. The coding frame helps the researcher to focus on select aspects of meaning. A systematic approach defines the steps needed to carry out the method. One approach is to perform a manual iterative process using an inductive and deductive qualitative coding technique to identify patterns and themes. An initial phase of open coding is used to organize the data and to become familiar with the themes. This open coding process will continue until no new information is gleaned. Then a process of selective coding will be used to define, develop, and refine the themes with attention to the nature and relationships of the categories and concepts emerging from the data [47, 59]. However, the approach and coding frame should have some flexibility in order to ensure that the coding frame matches the data being analyzed [59].

The steps of qualitative content analysis are formulating research question(s), preparing the data, constructing a coding frame, segmentation, trial coding, evaluating the coding frame, analysis, and interpreting the findings [59]. The first step is

formulating the research question(s). This helps the researcher focus on what are the salient concepts that wish to be gleaned from the data. The next step prepares the data for analysis. This includes the selection, collection, and converting of the data to a format that the chosen analysis software is able to read. The following step is to construct a coding frame. Schreier [59] describes the coding frame as the "heart of the method." The coding frame is comprised of at least one category that relates to the research question. Each category has at least two mutually exclusive subcategories that provide details about the concepts found in the data. Next segmentation of the data divides the data into the subcategories of the coding frame. The criteria for segmentation can be formal or thematic (Rustemeyer, 1992 via Schreier, 2014). Formal criteria are units such as words or sentences. A thematic criteria approach is able to overcome concepts that span the internal structure of the data and allows for tends to result in more meaningful units (Schreier, 2014). Once this step is complete, trial coding begins where the data is connected to the coding frame. The trial coding step allows the researcher to evaluate the coding frame. Once the coding frame is evaluated and potentially adjusted, the main analysis can now be completed. Finally, the method ends with the interpretation and presentation of the findings.

Semi-structured Interviews. Qualitative interviewing is a descriptive and interpretive research method that seeks meaning through the basic mode of conversation [39, 40]. Themes are introduced by the interviewer that is assumed to be of mutual interest to the participant. Interviewees respond with specific instances, examples of areas within the chosen themes. The interviewer and interviewee influence each other through the reciprocal act of conversation. Data is collected on the past, present, and/or perceived future phenomena of interviewees including “persons, events, activities, organizations, feelings, motivations, claims, concerns, … other entities” [43, p. 268]. These nuanced explorations provide researchers a way “to understand themes of the lived daily world from the [participants’] own perspectives” [40, p. 24] and provide the basis for interpretations of the described phenomena.

Qualitative interview research has little standardization in methodological conventions or rules allowing the specific purpose and topic to guide the investigation [39]. However, interviews may be categorized into two categories: semi-structured interviews and unstructured interviews [10]. Semi-structured interviews consist of a list of prepared questions that have a flexible order for allowances of follow-up and clarification questions. Interviewers play an active role in leading the interviews to ensure that the research questions are addressed. Unstructured interviews have no specific questions allowing the interviewee to direct the flow of conversation by introducing and structuring the problem in their own words corresponding to the broad issues raised by the interviewer. One strength of unstructured interviews is that they may be used as a discovery tool to avoid preconceived notions of the researcher.

Kvale [39] identifies twelve aspects of the interview form: life world, meaning, qualitative, descriptive, specificity, qualified naïveté, focus, ambiguity, change, sensitivity, interpersonal situation, and positive experience. (1) Lifeworld is the interviewee's everyday experiences and the meanings attached to these experiences that are expressed through an interviewee's own words. (2) The meaning of central themes is identified and interpreted by what is said, as well as, vocalization, facial expressions, and other bodily gestures. Interviews have a factual and meaning level. The factual level is concerned with what has occurred, while the meaning level focuses on how the person feels about an activity. (3) Qualitative knowledge is expressed in normal language with a precision in description and stringency in meaning interpretation. (4) Descriptions of experiences and feelings depict the differences and variations of the phenomenon. (5) Specificity of activities is of more important than general opinions. (6) Qualified naïveté should be exhibited by the interviewer. The interviewer should be open to unexpected phenomena and be critical of any pre-existing hypotheses during the interview. (7) The focus of an interview is on themes that are elicited from subjects through open-ended questions. (8) Ambiguity and contradiction can surface during an interview. It is important to clarify that the ambiguities are representative of the subject’s lifeworld and not a product of failed communication during the interview process. (9) During an interview, a participant may reflect on, and discover new aspects of, the themes being described. This will cause a change in the description and meaning of the themes provided by the participant. (10) Sensitivity towards a topic can influence an interviewer's statements on a theme. An interviewer's ability to obtain nuanced descriptions is influenced by their knowledge of the topic. (11) An interview is an interpersonal situation that produces knowledge in a reciprocally influenced interactions between the interviewer and interviewee. (12) A well-conducted interview can be a positive experience for the participant. An interviewer should be aware of anxiety-provoking dynamics within an interview.

Online Interviews. Although interviews are traditionally conducted in person, many studies exploring online communities may use computer-mediated communication such as online audiovisual media such as Skype and email [28]. Both communication modes provide the ability to interview people who are not easily accessible or geographically far apart. Online audiovisual media is the most similar to face-to-face interviews. Audiovisual media has the ability to detect non-verbal cues including facial expressions and body language, as well as vocal intonation of the participants.

Email is an asynchronous medium that may involve multiple e-mail exchanges over an extended period. Most of the studies addressing methodological issues occurred prior to 2003 suggesting that this method is a viable tool for qualitative research [46]. The advantages include the ability to interview people who are shy or have schedules that make it difficult to schedule an interview time [46], able to express themselves better in writing due to the interview being conducted in a second language [32] or provide interviewees more time to reflect on their responses [28]. Furthermore, email allows the possibility to conduct more than one interview at a time, reduced interviewer/interviewee effects such as visual or nonverbal cues

or status differences and increasing self-disclosure due to perceived anonymity of online communication. Although time is saved in transcription of the interviews due to already being in text form with fewer grammatical issues, however, there is a potential of interview data to be in multiple formats [28]. Challenges include data collection timeline is unpredictable from a week to several months, delays may be due to days or even weeks before a respondent replies to an e-mail, number of follow-up exchanges vary from 1 to 30 exchanges, limits the research to those people with access to the Internet, and unable to observe facial expressions and body language, or hear vocal intonation of the participants [28, 46]. Email interviews have similar recruitment response rate challenges of online surveys, however representative sample is not a goal in qualitative research. The issue of high non-delivery rate can be mitigated by inviting additional people to participate. Other challenges can be mitigated by substituting acronyms and emoticons for nonverbal cues and minimize participants' confusion by asking clear questions.

Procedures of Qualitative Content Analysis. The unit of analysis for this study is activity. The unit of observation for the content analysis is a talk page thread. These threads are in the Main Wikidata, WikiProject, user, and property talk pages. Analysis of these discussion threads aims to gain an understanding of how members participate in Wikidata; create items and properties in Wikidata; organize collaborations; and collectively detect, discuss, and resolve issues. The data used for content analysis is extracted from Main Wikidata, WikiProject, user, and property talk pages as well as, any open to the public official Wikimedia communications. These data are analyzed by a manual iterative process using an inductive open coding and deductive selective coding technique as discussed above to discover the relationships of the concepts in the data.

Procedures for Semi-structured Interviews. The unit of analysis for this study is activity. The unit of observation for the interviews is an individual Wikidata member. According to Wikidata, it has 17,600 editors who make at least one edit per month as of 2017. A subset of 8,000 active editors makes at least 5 edits per month. Subjects for interviews will be recruited from the open to the public Wikidata talk pages and official Wikimedia communications examined in the content analysis portion stage. A snowball approach will be used to identify subsequent participants. The number of interviewees suggested by Kvale and Brinkmann [40] is between 5 and 25. I hope to conduct twenty-four interviews. A semi-structured approach will be employed to allow for flexibility and alteration of the path of inquiry. Many studies exploring online communities use computer-mediated communication such as email, online audiovisual media (Skype and Google Plus Hangouts), and internet relay chat [28]. Thus, the interviews will be conducted using Skype or other similar communication software and a screen recorder will be employed as well as limited note-taking. Interviews will be transcribed and then coded using a grounded theory approach. Each interview will be transcribed and then coded using codes that arise from the interview itself. After each interview is coded, any new codes will then be applied to previous interviews, providing continuous evaluation of the content based on the information provided by the interview subjects.

The interview protocol was tested with five participants in order to discover any issues with conducting online interviews and refine the protocol itself. The protocol test participants comprised of men and women within an age range of 25-45 with occupations including librarian, government documents editor, and doctoral student. The process identified minor changes to the questions and supporting materials that provide more clarity.

5. CONCLUSION

An understanding of the activities in Wikidata, including the roles played, tools used, norms and rules followed, and solutions sought to address contradictions among different components of the activities will help inform communities wishing to contribute data to or reuse data from Wikidata. Furthermore, the findings of this study can go beyond understanding how Wikidata curates knowledge and potentially inform the design of other similar online production communities, scientific research institutional repositories, digital archives, and libraries. In conclusion, this research explores how editors participate in Wikidata and how they organize their work to add to the research of online collaborate communities.

REFERENCES

[1] L. Adamic, X. Wei, J. Yang, S. Gerrish, K. Nam, and G. Clarkson. 2010. Individual focus and knowledge contribution. *First Monday,* 15, 3.

[2] D. Anthony, S.W. Smith, and T. Williamson. 2009. Reputation and reliability in collective goods the case of the online encyclopedia Wikipedia. *Rationality and Society,* 21, 3, 283-306.

[3] J. Antin and C. Cheshire. 2010. Readers are not free-riders: Reading as a form of participation on Wikipedia. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing,*127-130. DOI: https://doi.org/10.1145/1718918.1718942

[4] J. Antin, C. Cheshire, and O. Nov. 2012. Technology-mediated contributions: Editing behaviors among new Wikipedians. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 373-382. DOI: https://doi.org/10.1145/2145204.2145264

[5] O. Arazy and O. Nov. 2010. Determinants of Wikipedia quality: The roles of global and local contribution inequality. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 233-236. DOI: https://doi.org/10.1145/1718918.1718963

[6] O. Arazy, O. Nov, R. Patterson, and L. Yeo. 2011. Information quality in Wikipedia: The effects of group composition and task conflict. *Journal of Management Information Systems,* 27, 4, 71-98. DOI: https://doi.org/10.2753/mis0742-1222270403

[7] O. Arazy, F. Ortega, O. Nov, L. Yeo, and A. Balila. 2015. Functional roles and career paths in Wikipedia. *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 1092-1105. DOI: https://doi.org/10.1145/2675133.2675257

[8] S. Ayvaz, J. Horn, O. Hassanzadeh, Q. Zhu, J. Stan, N.P. Tatonetti,.. . R.D. Boyce. 2015. Toward a complete dataset of drug-drug interaction information from publicly available sources. *Journal of Biomedical Informatics,* 55, 206-217.

[9] M. Balestra, C. Cheshire, O. Arazy, and O. Nov. 2017. Investigating the motivational paths of peer production newcomers. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'2017).*DOI: https://doi.org/10.1145/3025453.3026057

[10] K.M. Blee and V. Taylor. 2002. Semi-structured interviewing in social movement research. In B. Klandermans & S. Staggenbory (Eds.), *Methods of social movement research* (pp. 92-117). Minneapolis, MN: University of Minnesota Press.

[11] K. Carillo and C. Okoli. 2011. Generating quality open content: A functional group perspective based on the time, interaction, and performance theory. Information & Management, 48, 6, 208-219. DOI: https://doi.org/10.1016/j.im.2011.04.004

[12] J. Chen, Y. Ren, and J. Riedl. 2010. The effects of diversity on group productivity and member withdrawal in online volunteer groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 821-830. DOI: https://doi.org/10.1145/1753326.1753447

[13] B. Choi, K. Alexander, R. Kraut, and J. Levine. 2010. Socialization tactics in Wikipedia and their effects. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 107-116. DOI: https://doi.org/10.1145/1718918.1718940

[14] G. Ciampaglia and D. Taraborelli. 2015. MoodBar: Increasing new user retention in Wikipedia through lightweight socialization. *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*,734-742. DOI: https://doi.org/10.1145/2675133.2675181

[15] B. Collier and J. Bear. 2012. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 383-392. DOI: https://doi.org/10.1145/2145204.2145265

[16] P. Duguid. 2006. Limits of self-organization: Peer production and" laws of quality". *First Monday,* 11, 10.

[17] K. Ehmann, A. Large, and J. Beheshti. 2008. Collaboration in context: Comparing article evolution among subject disciplines in Wikipedia. *First Monday,*13, 10.

[18] A. Forte and A. Bruckman. 2005. Why do people write for Wikipedia? Incentives to contribute to open-content publishing. In *Group 2005 workshop: Sustaining community: The role and design of incentive mechanisms in online systems*. Sanibel Island, FL. Available at: http://www.cc.gatech.edu/aforte/ForteBruckmanWhyPeopleWrite.pdf.

[19] A. Forte, N. Kittur, V. Larco, H. Zhu, A. Bruckman, and R. Kraut. 2012. Coordination and beyond: Social functions of groups in open content production. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 417-426. DOI: https://doi.org/10.1145/2145204.2145270

[20] A. Forte, V. Larco, and A. Bruckman. 2009. Decentralization in Wikipedia governance. *Journal of Management Information Systems,* 26, 1, 49-72. DOI: https://doi.org/10.2753/MIS0742-1222260103

[21] R. Geiger and D. Ribes. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 117-126. DOI: https://doi.org/10.1145/1718918.1718941

[22] M. Gilbert, J. Morgan, D. McDonald, and M. Zachry. 2013. Managing complexity: Strategies for group awareness and coordinated action in Wikipedia. In *Proceedings of the 9th International Symposium on open collaboration*, 1-10. DOI: https://doi.org/10.1145/2491055.2491060

[23] A. Halfaker, R.S. Geiger, J.T. Morgan, and J. Riedl. 2012. The Rise and Decline of an Open Collaboration System: How Wikipedias Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist,* 57, 5, 664-688. DOI: https://doi.org/10.1177/0002764212469365

[24] A. Halfaker, A. Kittur, and J. Riedl. 2011. Don't bite the newbies: How reverts affect the quantity and quality of Wikipedia work. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11.

[25] R. Jesus, M. Schwartz, and S. Lehmann. 2009. Bipartite networks of Wikipedia's articles and authors: A meso-level approach. In *Proceedings of the 5th international symposium on wikis and open collaboration,* 5:1-5:10. DOI: https://doi.org/10.1145/1641309.1641318

[26] J. Jones. 2008. Patterns of Revision in Online Writing: A Study of Wikipedias Featured Articles. *Written Communication,* 25, 2, 262-289.

[27] V. Kaptelinin and B.A. Nardi. 2012. *Activity theory in HCI: Fundamentals and reflections*. Morgan & Claypool.

[28] M.M. Kazmer and B. Xie. 2008. Qualitative interviewing in Internet studies: Playing with the media, playing with the method. *Information, Communication, and Society,* 11, 257-278. DOI: https://doi.org/10.1080/13691180801946333

[29] B. Keegan and J. Brubaker. 2015. 'Is' to 'was': Coordination and commemoration in posthumous activity on Wikipedia biographies. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 533-546. DOI: https://doi.org/10.1145/2675133.2675238

[30] B. Keegan and D. Gergle. 2010. Egalitarians at the gate: One-sided gatekeeping practices in social media. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 131-134. DOI: https://doi.org/10.1145/1718918.1718943

[31] B. Keegan, D. Gergle, and N. Contractor. 2012. Do editors or articles drive collaboration?: multilevel statistical network analysis of Wikipedia coauthorship. In P*roceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, 427-436.

[32] B.S.K. Kim, B.R. Brenner, C.T.H. Liang, and P.A. Asay. 2003. A qualitative study of adaptation experiences of 1.5-generation Asian Americans. *Cultural Diversity & Ethnic Minority Psychology,* 9, 2, 156–170.

[33] A. Kittur and R. Kraut. 2008. Harnessing the wisdom of crowds in Wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 37-46.

[34] A. Kittur and R. Kraut. 2010. Beyond Wikipedia: Coordination and conflict in online production groups. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 215-224. DOI: https://doi.org/10.1145/1718918.1718959

[35] A. Kittur, B. Pendleton, and R. Kraut. 2009. Herding the cats: The influence of groups in coordinating peer production. *Proceedings of the 5th International Symposium on wikis and open collaboration*, 1-9. DOI: https://doi.org/10.1145/1641309.1641321

[36] R. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, ... J. Riedl. 2011. *Building successful online communities: Evidence-based social design*.Cambridge, Mass: MIT Press.

[37] M. Krieger, E.M. Stark, and S.R. Klemmer. 2009. Coordinating tasks on the commons: designing for personal goals, expertise and serendipity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1485-1494. DOI: https://doi.org/10.1145/1518701.1518927

[38] T. Kriplean, I. Beschastnikh, and D. McDonald. 2008. Articulations of WikiWork: Uncovering valued work in Wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on computer supported cooperative work*,47-56. DOI: https://doi.org/10.1145/1460563.1460573

[39] S. Kvale. 2007. *Doing interviews*.Los Angeles, CA: Sage. DOI: https://doi.org/10.4135/9781849208963

[40] S. Kvale and Brinkmann. 2009. InterViews: Learning the craft of qualitative research interviewing (2nd ed.). Thousand Oaks, CA: Sage.

[41] S.T.K. Lam, A. Uduwage, Z. Dong, S. Sen, D.R. Musicant, L. Terveen, and J. Riedl. 2011. WP:Clubhouse?: An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 1-10. DOI: https://doi.org/10.1145/2038558.2038560

[42] C. Lampe, J. Obar, E. Ozkaya, P. Zube, and A. Velasquez. 2012. Classroom Wikipedia participation effects on future intentions to contribute. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 403-406. DOI: https://doi.org/10.1145/2145204.2145267

[43] Y.S. Lincoln and E.G. Guba. (1985). Implementing the naturalistic inquiry. In *Naturalistic Inquiry* (pp. 250–288). Newbury Park, CA: Sage.

[44] J. Marlow and Dabbish. 2015. The effects of visualizing activity history on attitudes and behaviors in a peer production context. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 757-764. DOI: https://doi.org/10.1145/2675133.2675250

[45] P. Mayring. 2000. 'Qualitative content analysis', Forum Qualitative Sozialforschung/ Forum: Qualitative Social Research, 1, 2, Art. 20: http://nbn-resolving.de/urn:nbn:de:0114-fqs0002204.

[46] L. Meho. 2006. E-Mail interviewing in qualitative research: A methodological discussion. *Journal of The American Society for Information Science and Technology,* 57, 10, 1284-1295.

[47] A.J. Mills, G. Durepos, and E. Wiebe. 2010. *Encyclopedia of case study research*.SAGE, Thousand Oaks, CA.

[48] J.T. Morgan, M. Gilbert, M. Zachry, and D. McDonald. 2013. A content analysis of WikiProject discussions: Toward a typology of coordination language used by virtual teams. In *Proceedings of the 2013 conference on computer supported cooperative work companion,* 231-234. DOI: https://doi.org/10.1145/2441955.2442011

[49] J.T. Morgan, M. Gilbert, D. McDonald, and M. Zachry. 2014. Editing beyond articles: Diversity & dynamics of teamwork in open collaborations. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*,550-563. DOI: https://doi.org/10.1145/2531602.2531654

[50] O. Nov. 2007. What motivates Wikipedians? *Communications of the ACM,* 50, 11, 60-64.

[51] S. Oreg and O. Nov. 2008. Exploring Motivations for Contributing to Open Source: Initiatives: The Roles of Contribution Context and Personal Values. *Computers in Human Behavior,* 24, 5, 2055-2072.

[52] A. Piscopo, C. Phethean, and E. Simperl. 2017. Wikidatians are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base. In*50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*. AIS Electronic Library (AISeL).

[53] A. Piscopo, P. Vougiouklis, L. Kaffee, C. Phethean, J. Hare, and E. Simperl. 2017. What do Wikidata and Wikipedia have in common?: An Analysis of their Use of External References. In Proceedings of the*13th International Symposium on Open Collaboration (OpenSym '17)*. ACM, New York, NY, USA, Article 1, 10 pages.

[54] U. Pfeil, P. Zaphiris, and C.S. Ang. 2006. Cultural differences in collaborative authoring of Wikipedia. *Journal of Computer-mediated Communication,* 12, 1, 88-113.

[55] S. Ransbotham and G.C. Kane. 2011. Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly,* 35, 3, 613-627.
[56] L. Robert and D.M. Romero. 2015. Crowd size, diversity and performance. *Proceedings of the 33rd annual ACM conference on human factors in computing systems,* 1379-1382. DOI: https://doi.org/10.1145/2702123.2702469
[57] R.M. Ryan and E.L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist,*55, 1, 68-78.
[58] M. Schreier. 2012. *Qualitative Content Analysis in Practice*. London: Sage.
[59] M. Schreier. 2014. Qualitative content analysis. In Flick, U. *The SAGE handbook of qualitative data analysis* (pp. 170-183). London: SAGE Publications Ltd. DOI: https://doi.org/10.4135/9781446282243
[60] P. Shachaf and N. Hara. 2010. Beyond vandalism: Wikipedia trolls. *Journal of Information Science,* 36, 3, 357-370. DOI: https://doi.org/10.1177/0165551510365390
[61] G. Shao. 2009. Understanding the appeal of user-generated media: A uses and gratification perspective. *Internet Research,* 19, 1, 7-25. DOI: https://doi.org/10.1108/10662240910927795.
[62] J. Solomon and R. Wash. 2012. Bootstrapping wikis: Developing critical mass in a fledgling community by seeding content. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 261-264. DOI: https://doi.org/10.1145/2145204.2145247
[63] T. Steiner. 2014. Bots vs. Wikipedians, Anons vs. Logged-Ins (Redux): A Global Study of Edit Activity on Wikipedia and Wikidata. In *Proceedings of The International Symposium on Open Collaboration, OpenSym 2014, Berlin, Germany, August 27 - 29, 2014*. ACM, 25:1–25:7.
[64] B. Stvilia, M. Twidale, L.C. Smith, L. Gasser. 2008. Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology,*59, 6, 983-1001.
[65] B. Suh, G. Convertino, E.H. Chi, and P. Pirolli. 2009. The singularity is not near: Slowing growth of Wikipedia. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration (Article 8)*. New York, NY: Association for Computing Machinery.
[66] H. Tajfel. 1974. Social identity and intergroup behavior. *Social Science Information*. 13, 2, 65-93.
[67] H. Tajfel and J.C. Turner. 1979. An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel, *The social psychology of intergroup relations*. Monterey, CA: Brooks/Cole. pp. 33-47.
[68] T.P. Tanon, D. Vrandečić, S. Schaffert, T. Steiner, and L. Pintscher. 2016. From Freebase to Wikidata: The Great Migration. In *Proceedings of the 25th International Conference on World Wide Web* (WWW '16). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1419-1428.
[69] R. Tinati, M. Luczak-Roesch, N. Shadbolt, and W. Hall. 2015. Using WikiProjects to measure the health of Wikipedia. *Proceedings of the 24th international conference on world wide web,* 369-370. New York, NY, USA: ACM. DOI: https://doi.org/10.1145/2740908.2745937
[70] H. Ung and J. Dalle. 2010. Project management in the Wikipedia community. *Proceedings of the 6th International Symposium on wikis and open collaboration*, 1-4. DOI: https://doi.org/10.1145/1832772.1832790
[71] D. Vrandecic. 2013. The rise of Wikidata. *IEEE Intelligent Systems,* 28, 4, 90-95. DOI: https://doi.org/10.1109/MIS.2013.119
[72] L. Wang, J. Chen, Y. Ren, and J. Riedl. 2012. Searching for the goldilocks zone: Trade-offs in managing online volunteer groups. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 989-998. DOI: https://doi.org/10.1145/2145204.2145351
[73] M. Warncke-Wang, V. Ayukaev, B. Hecht, and L. Terveen. 2015. The success and failure of quality improvement projects in peer production communities. In *Proceedings of the 18th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 743-756. DOI: https://doi.org/10.1145/2675133.2675241
[74] R.K. Yin. 2018. *Case study research and applications: Design and methods* (5th ed.). Newbury Park, CA SAGE Publications, Inc.
[75] H. Zhu, R. Kraut, and A. Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW '12), 407-416. DOI: https://doi.org/10.1145/2145204.2145269
[76] H. Zhu, R. Kraut, and A. Kittur. 2012. Organizing without formal organization: Group identification, goal setting and social modeling in directing online production. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 935-944. DOI: https://doi.org/10.1145/2145204.2145344