Preliminary Exploration of Knowledge Curation Activities in Wikidata WikiProjects

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

ABSTRACT

Wikidata is one of the largest knowledge curation projects on the web. Their data is used by other Wikimedia projects such as Wikipedia, as well as major search engines. This qualitative study used content analysis of discussions involving data curation and negotiation in Wikidata. Activity Theory was used as a conceptual framework for data collection and analysis of the activities, members and tools. Some of the findings map Wikidata activities to curation frameworks. An understanding of the activities in Wikidata will help inform communities wishing to contribute data to or reuse data from Wikidata, as well as inform the design of other similar online peer-curation communities, scientific research institutional repositories, digital archives, and libraries.

KEYWORDS

Online curation communities; Curation; Wikidata

ACM Reference Format:
Timothy Kanke. 2018. Preliminary Exploration of Knowledge Curation Activities in Wikidata WikiProjects. In JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA. ACM, New York, NY, USA, 2 pages.
https://doi.org/10.1145/3197026.3203878

1 INTRODUCTION

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 [8]. According to Wikimedia February 2018 statistics, there are more than 50,000 contributors to Wikidata with an active base of 8,000 contributors making at least 5 edits per month. Wikidata is becoming the database of structured data for other information systems such as search engines. For example, one potential use is to provide updated information about clinical drug interactions (e.g., [1]). This paper reports on the early results of an exploratory study on how editors participate in Wikidata and how they organize their work.

A WikiProject is a group of individuals who want to collaborate to improve Wikidata on a specific topic or a task. Since little research has been conducted on WikiProjects in Wikidata, the following literature focuses on WikiProjects in a more well-documented Wikimedia project, Wikipedia. A reason to study WikiProjects is that they help editors to find others with similar topic interests or goals of improving Wikipedia by a maintenance task. These self-selected groups tend to have distinct activities and a shared identity. WikiProjects help support collaboration by focus attention towards coordinating and completing group related tasks [3], encourage good citizenship and maintenance behaviors [7]. WikiProject communications between editors may occur in additional pages outside of the project's talk page [4]. Some advanced editors seek additional information through talk pages of users of WikiProjects [5].

This study investigated the following research questions: (1) What are some of the curation and support activities being discussed by editors in Wikidata? (2) Who are the editors participating in these activities? (3) What tools are being used to perform these activities?

2 METHODOLOGY

To address the research questions, this study used content analysis of discussions involving data curation and negotiation in Wikidata. This study used Activity Theory as a methodological and conceptual framework for data collection and analysis with a focus on the activities, members, and tools. 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 [6].

The process of data collection and analysis was as follows. First, 181 Wikidata WikiProjects were identified. 153 projects focus on topics including astronomy, music, sports, and food. The remaining nineteen projects focus on knowledge bases and meta topics. From this set of WikiProjects, a smaller subset of four projects was selected: KOS, Ontology, Freebase, and Labels and Descriptions. The talk pages were analyzed by a manual iterative process using an inductive and deductive qualitative coding technique to identify patterns and themes. An initial

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

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203878

phase of open coding is used to organize the data followed by selective coding to define, develop, and refine the themes with attention to the relationships of the categories and concepts emerging from the data.

3 FINDINGS AND DISCUSSION

3.1 Activities

Many activities can be mapped to DCC Curation Lifecycle Model [2]. The DCC model uses the term "action" in a generalized sense while Activity Theory uses a more precise definition of action. Thus, this paper will replace "action" with "activities" to lessen confusion. Most of the activities that were identified occur in earlier stages of the DCC sequential activities. The activities are conceptualize, create or receive, appraise and select, and ingest. These activities support the goal of populating Wikidata with data from external sources such as freebase as well as original hack-a-thon content.

Several of the identified actions map to the DCC occasional activities of reappraising and dispose. Wikidata items are often created and then later expanded by many editors and bots by adding labels, descriptions in multiple languages, properties, and links to sources.

The other related activities that were identified are creating collaborative infrastructure, re-organizing collaborative infrastructure, and welcoming newcomers. Creating collaborative infrastructure include four actions: determining the scope of a project, creating work lists, determining the location of a list, and determining a location for discussion. As some WikiProjects grew in membership and data, it became necessary to reorganizing the collaborative infrastructure. For example, a talk page and table were divided into more manageable portions for each language. Welcoming newcomers is a critical issue for online communities because newcomers are a source of innovative ideas, work procedures, and resources. Actions that were identified focused on reorganizing help pages that have the potential to improve a new editor's experience.

3.2 Editors

True to the mission of Wikidata, the editors come from various Wikimedia projects including Wikipedia, Wikisource, and Wikitionary. Many editors also contribute to Wikipedia including the following languages: English, Russian, German, Dutch, Catalan, Macedonian, Oriya, Norwegian, Northern Sami, and Swedish. A few of the editors contribute to other Wikimedia projects: Wikisource and Wikitionary. Non-Wikimedia projects that were identified are OpenRefine, Schema.org, and Freebase. Some editors mention affiliation with the Wikimedia Foundation, a W3C working group, or Google. A few editors disclose their profession: librarian, professor, and software engineer.

3.3 Tools

Several tools were identified in the talk page discussions: internal communications, external hosted communications, and activity related tools. Wikimedia hosted communications include Talk pages in WikiProject, Property, Project chat, and User pages. Some editors use a WikiProject talk page to begin discussions before moving to the main project chat page. If a discussion that is relevant to a WikiProject is found to be occurring elsewhere, then there is a posting linking to that location to inform the project members. Occasionally editors use outside non-Wikimedia hosted platforms for communication and collaboration. Wikidata has a simplified user interface that lends itself to ease of contributing data, however, this simplicity may be at the cost of more advanced ontology creation tools. Some editors used Google Spreadsheet to collaborate on mapping an outside ontology to Wikidata, however, some editors expressed hesitations for collaborating on an outside resource due to privacy concerns. Bots are used to automate tasks including initial data population of properties and items from external sources. Scripts are explicitly mentioned less often than bots and are used to create constraint violation reports. Finally, Wikidata primary sources tool that is used in the process of ingesting donated data. Editors can review, edit, or reject data before importing the data.

4 CONCLUSION AND FUTURE WORK

The preliminary findings of this study identified the activities, participants, and tools used to curate knowledge. The identified activities are conceptualizing the curation process, appraising objects, ingesting objects from external sources, creating collaborative infrastructure, re-organizing collaborative infrastructure, and welcoming newcomers. Wikidata editors come from diverse backgrounds and are involved in other Wikimedia projects. Wikimedia hosts a majority of the communication, however, some WikiProjects need functionality or convenience that outside platforms provided. Future work will include expanding this study to investigate other aspects of Activity Theory including motivations, norms, rules, and roles of editors.

REFERENCES

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

[2] Digital Curation Centre. 2017. Curation Reference Manual. Retrieved from http://www.dcc.ac.uk/resources/curation-reference-manual

[3] A. Forte, N. Kittur, V. Larco, H. Zhu, A. Bruckman, and R. E. Kraut. 2012. Coordination and beyond: Social functions of groups in open content production. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, 417-426.

[4] M. Gilbert, J. T. Morgan, D. W. 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, ACM, 5:1-5:10.

[5] H. Jung, S. R. Hong, P. Meas, and M. Zachry. 2015. Designing tools to support advanced users in new forms of social media interaction. In Proceedings of the 33rd Annual International Conference on the Design of Communication, ACM, 34:1-34:10.

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

[7] A. Kittur, B. Pendleton, and R. E. Kraut. 2009. Herding the cats: The influence of groups in coordinating peer production. In Proceedings of the 5th International Symposium on Wikis and Open Collaboration, ACM, 7:1-7:9.

[8] D. Vrandecic and M. Krtotzsch. 2014. Wikidata: A free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.