

Deliberative Quality in Wikimedia Projects: Comparing Requests for Comments on Meta, Wikipedia and Wikidata

Dylan Baumgartner
University of Zurich
dylan.baumgartner@uzh.ch

Cristina Sarasua
University of Zurich
sarasua@ifi.uzh.ch

Pablo Aragón
Wikimedia Foundation
paragon@wikimedia.org

Abstract

Requests for Comments (RfCs) provide a central deliberation space in Wikimedia projects. Analyzing the deliberative quality of RfCs discussions is essential because high-quality deliberation can lead to better decision-making. In this extended abstract, we present ongoing work focusing on comparing deliberative quality in *Meta*, the *English Wikipedia* and *Wikidata*.

Keywords: RfCs, Deliberation quality, Wikipedia, Wikidata, Meta, Group diversity

Introduction

The Wikimedia community relies on deliberation as a key instrument for self-governance in projects like Wikipedia and Wikidata (Black et al., 2011). Decisions in Wikimedia projects are expected to be made primarily by consensus, with disputes between editors addressed through respectful and civilized dialogue. Requests for Comments (RfCs) provide a dedicated area to discuss topics that require the attention and opinion of a large portion of the community.

Political science scholars have extensively studied the notion of discourse quality in the context of deliberative democracy. There is a full body of work around the *Discourse Quality Index (DQI)* (Steenbergen et al., 2003) — a coding scheme to label the discourse quality of political debates based on multiple dimensions such as respect, as well as level and content of justification. The scheme initially relied on Habermas' discourse ethics and was later extended beyond rational deliberation on a second version (DQI2.0) (Bächtiger et al., 2010). We use the DQI as a conceptual framework to define deliberative quality.

Our goal is two-fold: (i) to compare the deliberative quality of discussions taking place in RfCs of the English Wikipedia (the free encyclopedia), Wikidata (the free knowledge base), and Meta (the website for the global Wikimedia community); and (ii) to study the relationship between group diversity and deliberative quality.

Related Work

Most studies on deliberation on Wikimedia have focused on article talk pages of English Wikipedia. Research

has shown that discussions there serve multiple purposes (Viégas et al., 2007), which might vary by the type of article (Schneider et al., 2010). A common purpose is the enforcement of Wikipedia policies, given that ambiguities in their formulation have given rise to power plays in article talk page discussions (Kriplean et al., 2007).

Several coding schemes for article page talk discussions have been created to identify dialog acts (Ferschke et al., 2012), dispute tactics (de Kock and Vlachos, 2022), and even deliberative argumentation strategies (Al Khatib et al., 2018). Furthermore, similar works analyzing deliberation of discussions on Articles for Deletion pages (Xiao and Askin, 2014) and RfCs (Im et al., 2018) also developed their own coding strategies. Hence, research on the deliberative quality of Wikipedia discussions has barely exploited state-of-the-art approaches to measure political deliberation (e.g., DQI).

Data

We collected a dataset of closed RfCs using pywikibot¹ looking for pages referencing specific templates indicating the closure of an RfC. Different projects, follow different practices: Wikipedia RfCs are embedded in talk pages. So, to identify closed RfCs, we looked for instantiations of `{{closed rfc top}}` and parsed the content until `{{closed rfc bottom}}`. Wikidata and Meta RfCs, in contrast, have separate pages. Hence, we looked for the occurrence of `{{RfCSubpage|closed=yes}}` in Wikidata pages and the occurrence of `{{rfc subpage}}` in Meta pages that had the status parameter set to one of the values in the following list: "resolved", "closed", "globalban=yes", "globalban=no".

We parsed the Wikitext into individual comments preserving the original indentation structure using the Grawitas parser². Moreover, we retrieved public editor information from the ten largest Wikipedia language versions, MediaWiki, Wikidata and Meta replica databases available on Toolforge³. To keep a human-readable version of

¹<https://github.com/wikimedia/pywikibot>

²<https://github.com/bencabrera/grawitas>

³<https://admin.toolforge.org/>

the content, we used the `mwparsersfromhell` parser⁴ and obtained the labels of templates using the MediaWiki API⁵.

In total, our dataset comprises 2,302 individual RfCs, out of which 1,552 RfCs are from Wikipedia, 140 from Wikidata, and 610 from Meta. Across these RfCs, we collected 92,221 comments from 8,395 distinct editors.

Methodology

We measure the deliberative quality of RfCs by applying the DQI framework. For each comment that an editor added to the RfC, we measure the presence or lack of a selection of the DQI2.0 dimensions that are relevant for our data (i.e., we do not consider *interruption* because, in our scenario, discussions are held online in an asynchronous way; we also exclude deliberative negotiation). Specifically, we focus on the following dimensions: *explanation*, *causal reasoning*, *advocacy*, *public interest*, *counterargument*, *respect*, *disrespect*, *question*, *response*, *constructive proposal*, and *narrative*. Figure 1 shows an example of the binary labels that we assign to a comment present in a Wikidata RfC. The text contains an explicit question, and the editor explains why it would be negative not to be able to use items as sources. Hence, we label *explanation*, *causal reasoning*, and *question* as 1, and the rest of the dimensions as 0.

We are currently labeling a subset of 500 comments randomly sampled from our dataset. Then, we will scale the labeling process using the supervised machine learning approach to classify speech according to the DQI2.0 implemented in (Fournier-Tombs and MacKenzie, 2021).

Group diversity will be measured by applying the widely-used diversity metric Shannon-Entropy (Shannon, 2001) on public features of editors (e.g., edit count, role, age, activity across projects).

We plan to run a correlation study between these variables, and for the comparative analysis across projects, we plan to use both visual descriptive statistics, as well as comparative hypothesis testing.

References

- [Al Khatib et al.2018] Khalid Al Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555.
- [Bächtiger et al.2010] André Bächtiger, Susumu Shikano, Seraina Pedrini, and Mirjam Rysler. 2010. Measuring deliberation 2.0: Standards, discourse types, and sequentialization.
- [Black et al.2011] Laura W Black, Howard T Welsler, Dan Cosley, and Jocelyn M DeGroot. 2011. Self-governance through group discussion in wikipedia: Measuring deliberation in online groups. *Small Group Research*, 42(5):595–634.
- [de Kock and Vlachos2022] Christine de Kock and Andreas Vlachos. 2022. How to disagree well: Investigating the dispute tactics used on wikipedia. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, pages 3824–3837. ACL.
- [Ferschke et al.2012] Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar. 2012. Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786.
- [Fournier-Tombs and MacKenzie2021] Eleonore Fournier-Tombs and Michael K MacKenzie. 2021. Big data and democratic speech: Predicting deliberative quality using machine learning techniques. *Methodological Innovations*, 14(2):20597991211010416.
- [Im et al.2018] Jane Im, Amy X. Zhang, Christopher J. Schilling, and David Karger. 2018. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceeding of CSCW 2018*, nov.
- [Kriplean et al.2007] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: Cs*w or how policy mediates mass participation. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work, GROUP '07*, page 167–176, New York, NY, USA. ACM.
- [Schneider et al.2010] Jodi Schneider, Alexandre Passant, and John G Breslin. 2010. A content analysis: How wikipedia talk pages are used. In *Proceedings of Web Science 2010*, pages 1–7.
- [Shannon2001] Claude Elwood Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE*, 5(1):3–55.
- [Steenbergen et al.2003] Marco R Steenbergen, André Bächtiger, Markus Spöndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.
- [Viégas et al.2007] Fernanda B Viégas, Martin Wattenberg, Jesse Kriss, and Frank Van Ham. 2007. Talk before you type: Coordination in wikipedia. In *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pages 78–78. IEEE.
- [Xiao and Askin2014] Lu Xiao and Nicole Askin. 2014. What influences online deliberation? a wikipedia study. *Journal of the Association for Information Science and Technology*, 65(5):898–910.

⁴<https://github.com/earwig/mwparsersfromhell>

⁵<https://www.mediawiki.org/wiki/API:Expandtemplates>

“*Does that mean, that an item can't be used as a source too? That would be disadvantageous an lead to doubled information since there are already many items about books and websites.”

DIMENSION	LABEL
Explanation	1
Causal reasoning	1
Advocacy	0
Public Interest	0
Counterargument	0
Respect	0
Disrespect	0
Question	1
Response	0
Constructive proposal	0
Narrative	0
Deliberative negotiation	0

Figure 1: Labeling of a Wikidata RfC comment using the 12 DQI2.0 dimensions selected in our methodology.
Source: https://www.wikidata.org/wiki/Wikidata:Requests_for_comment/References_and_sources