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 (Viegas 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 Vla- chos, 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 pywikibot 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 parser. Moreover, we retrieved public editor information from the ten largest Wikipedia language versions, MediaWiki, Wikidata and Meta replica databases available on Toolforge. To keep a human-readable version of
the content, we used the mwparserfromhell package and obtained the labels of templates using the MediaWiki API. 

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


“*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.”

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>LABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanation</td>
<td>1</td>
</tr>
<tr>
<td>Causal reasoning</td>
<td>1</td>
</tr>
<tr>
<td>Advocacy</td>
<td>0</td>
</tr>
<tr>
<td>Public Interest</td>
<td>0</td>
</tr>
<tr>
<td>Counterargument</td>
<td>0</td>
</tr>
<tr>
<td>Respect</td>
<td>0</td>
</tr>
<tr>
<td>Disrespect</td>
<td>0</td>
</tr>
<tr>
<td>Question</td>
<td>1</td>
</tr>
<tr>
<td>Response</td>
<td>0</td>
</tr>
<tr>
<td>Constructive proposal</td>
<td>0</td>
</tr>
<tr>
<td>Narrative</td>
<td>0</td>
</tr>
<tr>
<td>Deliberative negotiation</td>
<td>0</td>
</tr>
</tbody>
</table>