What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata

Alessandro Piscopo, Chris Phethean, Elena Simperl

University of Southampton,
Southampton, United Kingdom
{A.Piscopo, C.J.Phethean, E.Simperl}@soton.ac.uk

Abstract. Wikidata is a community-driven knowledge graph which has drawn much attention from researchers and practitioners since its inception in 2012. The large user pool behind this project has been able to produce information spanning over several domains, which is openly released and can be reused to feed any information-based application. Collaborative production processes in Wikidata have not yet been explored. Understanding them is key to prevent potentially harmful community dynamics and ensure the sustainability of the project in the long run. We performed a regression analysis to investigate how the contribution of different types of users, i.e. bots and human editors, registered or anonymous, influences outcome quality in Wikidata. Moreover, we looked at the effects of tenure and interest diversity among registered users. Our findings show that a balanced contribution of bots and human editors positively influence outcome quality, whereas higher numbers of anonymous edits may hinder performance. Tenure and interest diversity within groups also lead to higher quality. These results may be helpful to identify and address groups that are likely to underperform in Wikidata. Further work should analyse in detail the respective contributions of bots and registered users.

Keywords: Wikidata, collaborative knowledge graphs, group composition

1 Introduction

Peer production systems have been experimented with successfully in several fields. Wikipedia is probably the most well-known example, but the efforts of communities of users are behind diverse projects, including open source software (e.g. Apache or Linux), database management systems (e.g. PostgreSQL), or question-answer sites (e.g. Stack Overflow). Wikidata is a recent addition to this already large list. It is a community-driven knowledge graph started by the Wikimedia foundation in October 2012. Since its inception, it has gathered a user pool of around 100 thousand registered users, who are able to add facts about more than 24 million entities. Because of these and other features, Wikidata has drawn the attention of researchers and practitioners alike.

Knowledge Graphs (KGs) are large collections of structured data, encoded as terms describing entities and the relationships existing between them [26]. KGs are important as they provide data that can be processed by machines to create new, tailored information. For example, Wikidata was initially designed as a structured backbone for Wikipedia. Its primary aim was to offer an improved model to maintain the structured knowledge already contained in Wikipedia and make it available to other applications online. Besides, the use of a central store of knowledge and facts allows each localised version of Wikipedia to access the same data for each page or topic, improving coverage across the different languages of the site. Wikidata's aims go beyond the support of other Wikimedia projects. Its data is released under an open licence which grants free reuse and sharing. Hence, it can act as a source for any information-based application and, due to the large number of identifiers from and links to other resources, help integrate knowledge from several sources.

Wikidata is reliant on its community for adding and maintaining data. Its contributors can be divided into three types: bots (software programmed to perform edits and maintenance work) and registered and anonymous human editors. Bots carry out a large share of work in Wikidata [27]. However, although their activities have been analysed with regard to the type of tasks they perform on the overall KG [21], it is not yet clear to what extent they contribute to the quality of Wikidata.

Concerning human editors, the openness and versatility of the wiki model [32] allows users to collectively author structured information, with no official editorial oversight. People of different backgrounds, skills, and perspectives put their efforts together to build Wikidata's knowledge. An extensive body of literature analyses the effects of group composition and diversity, which we define as the distribution of members in a group with respect to a common feature [12]. Findings for off– and online contexts show that differences within a group may be a double-edged sword when it comes to performance. Analysing the effects of group demography on team outcomes, Ancona & Caldwell [2] show that diversity negatively influences performance, both directly and mediated by internal processes and communication. On the other hand, Jehn *et al.* [14] found that various types of diversity are positively related to perceived (social diversity) and actual group performance (informational diversity). Concerning online contexts, more heterogeneous groups are more likely to take better decisions [16] and, when the level of conflict is high, to create higher quality articles in Wikipedia [3]. To the best of our knowledge, no study has yet addressed the connections between group composition and diversity and the resulting outcome quality in Wikidata.

The current study aims to address the above mentioned gaps by investigating the 'right mix' of users that leads to good quality in Wikidata. First, it analyses the relationship between the share of contributions of bots, registered, and anonymous human users and outcome quality. Second, it investigates the effect of the distribution of two features within Wikidata groups: length of activity within the community and task knowledge, i.e. tenure and interest diversity. These two variables were chosen due to the variations in activity across different levels of experience and different tasks reported by Wikidata editors in previous work [24].

Gaining insights about the relationship between group composition and performance is essential to improve the understanding of how Wikidata's knowledge is created and of the underpinning quality-related processes. It would enable one to recognise the community patterns that lead to good performance in the system, and possibly intervene on those which are likely to cause quality issues. Wikidata has been defined as a combination between a peer-production system and a collaborative ontology project [21]. Additionally, the size of its user pool has not been attained by previous analogue projects, such as Freebase [11]. These features set Wikidata apart from several online projects examined by the literature over group composition. Hence, this study extends previous observations about the effects of group diversity on the performance of online communities to collaborative KGs.

In the next section we present various aspects of Wikidata. Subsequently, we review a selection of works of relevance about group diversity and outline the research hypotheses tested in the current study. Following, we present methods and data employed. Finally, we provide results and discuss these and the limitations of our work.

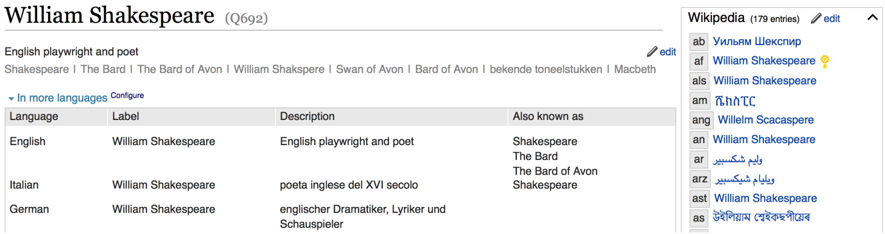
2 Wikidata

Wikidata aims to create a general domain KG maintained by a community of users, openly shareable and reusable, and accessible by machines to perform reasoning and provide complex answers to queries. The next sections discuss its features in detail.

2.1 The Data Model of Wikidata

Knowledge in Wikidata is expressed through *items* and *properties*. Items refer to concrete things or characters (e.g. the Colosseum or William Shakespeare) or to abstract concepts and categories, e.g. humans or music. Properties define relationships between items and other items, or between items and literal values, and are used to state facts.

Language-independent URIs are used to refer to items and properties, and editors can add human-readable *labels*, *descriptions*, or *aliases* in up to 358 languages (see Figure 1). This feature allows any user to edit each entity, regardless of his/her language. Structured data that goes beyond these simple labels and descriptions is represented using relationships in the form of property-value pairs named *claims*. These are the base of the data model of Wikidata. A claim may include an optional *reference*, (a link to its source) and/or a *qualifier* to provide additional contextual information for the claim, such as specifying the date that it refers to (Figure 2(a)). Both items and properties are defined through statements, and a number of labels, aliases, and descriptions. Properties specify a permitted value datatype, e.g. a number (for a property such as 'population')



William Shakespeare (Q692)

English playwright and poet

Shakespeare | The Bard | The Bard of Avon | William Shakspere | Swan of Avon | Bard of Avon | bekende toneelstukken | Macbeth

▾ In more languages Configure

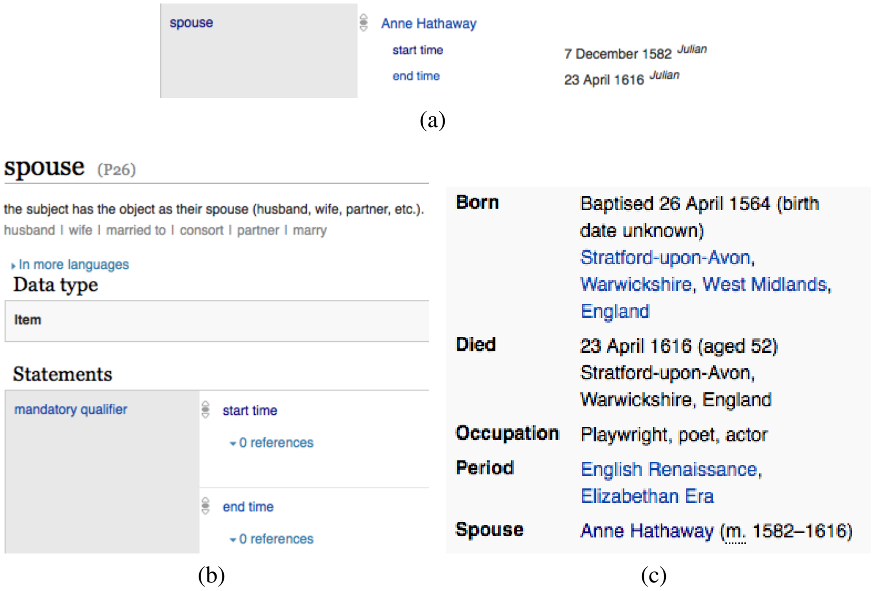
Language	Label	Description	Also known as
English	William Shakespeare	English playwright and poet	Shakespeare The Bard The Bard of Avon
Italian	William Shakespeare	poeta inglese del XVI secolo	Shakespeare
German	William Shakespeare	englischer Dramatiker, Lyriker und Schauspieler	

Wikipedia (179 entries)

ab	Уильям Шекспир
af	William Shakespeare
als	William Shakespeare
am	ሼክስፒር
ang	Willelm Scacaspere
an	William Shakespeare
ar	وليم شكسبير
arz	ويليام شيكسبير
ast	William Shakespeare
as	উইলিয়াম শ্বেইকছপীয়েৰ

Fig. 1. Multilingual labels, description, and aliases (under "Also known as") in a Wikidata Item. On the right, links to all Wikipedia language versions of the related article.

or another Wikidata item (for a property such as ‘spouse’ — see Figure 2). Additionally, they may include constraints, i.e. conditions that have to be met when the property is used. Constraints span from defining the domain and range to other characteristics of the property, e.g. requiring a symmetric property. However, constraints are not enforced in Wikidata’s data model [9], and are used only for quality control purposes.



spouse	Anne Hathaway	
	start time	7 December 1582 Julian
	end time	23 April 1616 Julian

(a)

spouse (P26)

the subject has the object as their spouse (husband, wife, partner, etc.).
husband | wife | married to | consort | partner | marry

▸ In more languages

Data type

Item

Statements

mandatory qualifier	start time ▾ 0 references
	end time ▾ 0 references

(b)

Born	Baptised 26 April 1564 (birth date unknown) Stratford-upon-Avon, Warwickshire, West Midlands, England
Died	23 April 1616 (aged 52) Stratford-upon-Avon, Warwickshire, England
Occupation	Playwright, poet, actor
Period	English Renaissance, Elizabethan Era
Spouse	Anne Hathaway (m. 1582–1616)

(c)

Fig. 2. Example Property for Shakespeare: ‘Anne Hathaway’ is another item, therefore the property in (a) states that Shakespeare (item) has a ‘spouse’ (property) which is ‘Anne Hathaway’ (another item). ’Spouse’ (b) specifies qualifiers to provide additional details to the claim. On Wikipedia this relationship is displayed in an InfoBox (c).

2.2 Quality in Wikidata

The body of research around data quality in Wikidata has grown in recent years. Inconsistencies in Wikidata’s taxonomy hierarchies are the focus of [7], which found many erroneous patterns deriving from misuse of properties such as *instance of* or *subclass*. Other research has compared the quality of data in Wikidata to other KGs. [29] studied Wikidata and DBpedia with respect to their fitness for open-domain question answering. Wikidata outperforms DBpedia in several of the domains tested. Fa̋rber *et al.* [10] evaluate Wikidata and four other KGs (DBpedia, Freebase, YAGO, and OpenCyc) along several data quality dimensions. Wikidata reaches quality levels comparable to the other KGs examined, if not higher in some cases, e.g. schema completeness.

The approaches mentioned above measured data quality over the whole KG or a subset of it. The current study focuses on the effects of group diversity on outcome quality, thus we sought a quality measure of elements identifiable as the result of the work of a group of editors. Items are the outcome of the cooperation of a set of users—a group—and are the building blocks of Wikidata's knowledge. For our analysis we used the single-grading scale quality labels in [33]. The advantages of this measure are twofold. First, it provides an overall score for item quality. This feature was appropriate for our aim to investigate how group diversity relates to performance in general, rather than to specific aspects of quality. Second, the item quality labels are based on criteria set up by the Wikidata community and are generated by the community itself. Hence, it is a meaningful measure of quality, as perception of information contributors is key to define quality in user-contributed systems [18]. We provide further details about this quality measure in section 4.1.

2.3 Editing Wikidata

The community is responsible for adding and maintaining all the elements that constitute Wikidata. While the majority of these operations are very simple, requiring little skills or knowledge other than factual information, other tasks are more demanding and can potentially influence a larger part of the KG. These differences afford the classification of tasks into lightweight and heavyweight, according to the definition given in [13]. As an example of a lightweight task, if the claim about Shakespeare's place of birth was wrong, users could modify it, linking the *place of birth* property to the *Stratford-upon-Avon* item. The same applies to labels: if users want to include the Italian name for London, they only need to add it in the appropriate place. Adding and modifying claims and labels requires little specialised knowledge and is largely independent from other revisions; these tasks account for the vast majority of Wikidata edits and can be classified as lightweight contributions. Other less trivial tasks can be defined as heavyweight. Editing statements using the properties *instance of* or *subclass of* requires to be familiar enough with knowledge engineering concepts to understand the difference between the two and use them accordingly. The same skills are required to edit properties, whereby understanding their relationship with other properties, the constraints that may be applied to them, and their intended use is essential. The item-building process may thus be decomposed into several smaller tasks, which are loosely dependent from each other and include both light– and heavyweight tasks.

2.4 The Wikidata Community

Wikidata editors can be either bots or humans. Since the Wikidata philosophy allows anyone to contribute, human editors can either register or contribute anonymously.

Bots are developed by users and should be approved by the community before performing any operation[1]. Notwithstanding, several bots are functioning without a previous formal approval. Each bot is operated by a human user, who should maintain it and ensure that it does not cause any damage to the KG.

[1] The official Wikidata policy about bots is in `https://www.wikidata.org/wiki/Wikidata:Bots`, which is cited throughout this section. This was also the source for our list of bots and active users, together with `https://www.wikidata.org/wiki/Category:Bots_without_botflag` and `https://www.wikidata.org/wiki/Special:Statistics`

Although their number is small compared to human users (399 vs. around $17,000$ monthly active users), bots do the lion's share of Wikidata edits ($> 95\%$) [27]. Almost 90% of bots' editing activities concerns the addition or modification of item statements (58%) or labels/descriptions/aliases (30%) [21]. According to the Wikidata policy about bots, these are required to include a reference with every new statement and to check if any revision they make violates any property constraints. Regarding the scope of bot edits, they often focus on a class or typology of items or statement. As an example, some bots add statements to items in the biomedical field, others focus on settlements of determined countries, and others map Wikidata items to their equivalent in other KGs, such as DBpedia. Human edits can override bot edits, and vice versa. No preference is granted to any user type over the other. Besides editing, bots perform a variety of maintenance tasks, which include checking property constraint violations, moving pages, or fixing wrong user names. Moreover, as Wikidata emerged, bots were created to move language links from Wikipedia over to Wikidata, in order to help build inter-language links connecting different language versions of Wikipedia pages [31].

Finally, human users can contribute anonymously. The revisions authored by these users are a small percentage of all Wikidata edits (0.5%), according to statistics we extracted about Wikidata editing activity. Anonymous users have been studied less since their activity is difficult to track over time. However, they contribute to properties in higher measure than bots.

3 Group Diversity and Outcome Quality

The effects of group diversity on outcome quality have been thoroughly investigated, both in offline and online contexts. Diversity appears to be both an opportunity as well as a challenge for work teams [19]. Differences among group members may generate a "creative abrasion" that positively affects performance [3]. On the other hand, diversity may hamper the identification of users within a group [19]. Researchers have tried to explain these mixed effects by categorising various types of diversity. In their review of studies about diversity in organisational groups, Milliken & Martins [19] distinguish observable and underlying attributes. Dissimilarity with regard to observable attributes, such as age, gender, or race, lead to higher turnover and lower integration [19]. Underlying attributes may refer to personality characteristics, values, skills and knowledge, and functional background, among others. Skills– or knowledge-related diversity affects positively performance in top-management and project teams, whereas the effects of other types of underlying attributes are less clear. In a similar fashion, Arazy *et al.* [3] identify surface– and deep-level diversity. Whereas the first encompasses demographic characteristics, the latter regards expertise, knowledge, and functional background. Deep-level diversity entails a higher variety of perspectives, which create better conditions for creativity and knowledge sharing. Other attempts to interpret different types of diversity see two contrasting viewpoints about its effects on group performance, the *social category* and the *information/decision making* perspectives [30]. The social category perspective focuses more on relational aspects. Homogeneous groups benefit from higher cohesion and member commitment, thus being able to produce a better output. The information/decision making perspective is shifted more towards job-related attributes and connected to less evident aspects of members, e.g. educational or functional background. Diversity also influences positively performance according to this perspective. In [30] these two perspectives are combined, by connecting them to the requirements and the elaboration of tasks. Diversity would lead to better performance in case of complex information-processing tasks, with respect to simple, repetitive ones.

A great deal of previous research has focused on demographic features of group members. [2] explored direct and indirect effects on performance by the distribution of organisation tenure and functional speciality in the team, with mixed results. Whereas both types of diversity have a direct negative effect on team– and managerial–rated performance, their indirect effects look more complex. More heterogeneous groups with regard to tenure, i.e. the length of activity within a team, are able to define better their goals and priorities. Higher functional diversity improves external communication. Both clarity of goals and priorities and improved external communication positively affect performance. These conflicting findings suggest a complex relationship between group diversity and outcomes, with effects that may change according to the context and the type of diversity studied. Pelled *et al.* in [23] draw similar conclusions about the complexity of the relationship between several types of diversity, conflict, and performance. Their study, carried out on corporation teams, finds a positive association between tenure and functional diversity and task conflict. In turn, this affects positively cognitive task performance, thus suggesting that differences in organisational tenure and functional background of group members may indirectly improve their outcome. Other variables, e.g. race and gender, do not seem to directly influence conflict.

With regard to online systems, [3,8,16] analyse how diversity affects outcome quality in Wikipedia, obtaining similar results. The relationship between cognitive diversity, task conflicts, and quality of Wikpedia articles is analysed in [3]. Cognitive diversity refers to the mental models and interests of the members of a group and positively influences outcome performance. The effects of tenure diversity on the quality of the decisions to delete Wikipedia articles are analysed in [16]. Whereas the presence of newcomers by itself appears detrimental for outcome quality, in agreement with previous literature on offline settings [20], a moderate tenure diversity is related to higher quality decisions. Chen *et al.* [8] study how interest and tenure diversity influence productivity and withdrawal in Wikipedia projects. Interest diversity is a concept close to cognitive and functional diversity. It refers to the variety of members' interests in a group. In collaborative projects such as Wikipedia or Wikidata, where users contribute voluntarily and generally choose which tasks to take on, an individual's interests may actually determine their activity and function within the project. According to [8], tenure diversity leads to higher productivity, but with diminishing results, while increasing member withdrawal, analogously to what noted in [16]. Interest diversity is linearly correlated to productivity, whereas no evidence is found about its influence on member withdrawal.

The current study focuses on tenure and interest diversity on Wikidata. Concerning tenure, previous work [24] has shown that Wikidata editors change the focus and scope of their activity as they gain experience within the system. This suggests that users with different levels of experience may bring complementary skills in building high-quality items. Editor activity may also vary along their edit scope [24]. Some users carry out similar tasks, i.e. adding references, on a broad spectrum of items, whereas others focus on a restricted number of items, specialising on a single domain. The contribution of these two types of users may thus be necessary to create good quality items.

3.1 Research Hypotheses

This section presents the hypotheses tested in this study. Hypotheses 1-3 concern the proportion of contributions by different user types. Hypotheses 4-5 regard tenure and interest diversity.

The importance of bot contributions for outcome quality has been noted already with regard to Wikipedia [22]. In Wikidata, the amount of bot editing activity and its scope means that their contribution is a crucial factor for outcome quality. Therefore, we formulate our first hypothesis:

Hypothesis 1: The percentage of bot edits is positively related to item quality.

Although the contribution of bots is important to set the basic structure of items—e.g. automatically adding Wikimedia links and labels in several languages —some tasks require human editors. These possess the knowledge and skills to add descriptions and aliases, and perform quality controls that are not routinely performed by bots. In their analysis of the emergence of user roles connected to the division of labour in Wikidata, Müller-Birn *et al.* observe that bots and humans perform similar tasks, however with a different distribution [21]. Bots' activities focus more on setting new statements, whereas human contributors primarily edit them and add references. Hence, bots and registered human editors may need to complement their effort in order to achieve high-quality items. On the other hand, users who edit anonymously may have lower levels of attachment and have shown to often generate spam and vandalism in other projects [1]. We refer to interaction between human and bot editors as the balance of the respective contributions to an item. Higher interaction means a more equal contribution from each of these two user types.

Hypothesis 2: High levels of interaction between human and bot users are positively related to quality.

Hypothesis 3: The percentage of anonymous human edits is negatively related to item quality.

As mentioned above, Wikidata editors take on different types of tasks along their evolution as part of the community [24]. Seasoned users focus on higher-level tasks, e.g. working on the conceptual structure of knowledge and on quality maintenance tasks, whereas newcomers tend to concentrate their efforts on adding and modifying statements. Items edited by users with various tenure levels may benefit from these different "specialisations". Additionally, more experienced users feel a sense of responsibility towards Wikidata. This might drive them to oversee the work done by other editors and help ensure quality.

Hypothesis 4: Tenure diversity is positively related to item quality.

A similar process may be at play with regard to interest diversity. Editors working on a broader range of items may lead to different perspectives to the creation of items. One of the peculiarities of KGs is that the entities they contain are linked, allowing machines to perform inferences and reason following these connections. Users with heterogeneous interests may facilitate the creation of internal links.

Hypothesis 5: Interest diversity is positively related to item quality.

4 Data and Methods

We describe in the following the approach employed to test our research hypotheses, including the variables examined, the analysis strategy, and the data used.

4.1 Dependent Variable

For the purpose of this study, we used a measure of quality generated by the authors of [33] in close-collaboration with the community of Wikidata. It is a single-grading scale which assigns labels to items from A (the highest) to E. The criteria on which the scale is based comprehend the completeness of the item, seen as the number of relevant statements; the plurality of the sources used to support the statements; labels and descriptions in an appropriate number of languages; links to other wiki projects; and possibly whether media files are attached[2]. Quality criteria were reviewed through discussions with the community, both on– and offline.

Item labels were collected for a sample of $5,000$ Wikidata items, each evaluated by a Wikidata editor. A pilot campaign was previously run to verify and refine the quality of community-generated labels. The sample selection aimed to obtain a more balanced distribution of item quality classes, compared to the entirety of Wikidata, where the majority of items likely fall in classes C to E. Therefore, the authors of [33] over-represented classes A and B by selecting a certain number of items per size (in bytes), following the assumption that larger items would more likely have higher quality. Additionally, they included a number of 'special items', i.e. items whose ID has a particular meaning. The distribution of items per quality level is in Table 1.

Table 1. Distribution of quality levels.

Quality level	No. items	No. items (w/ at least 1 human edit)
A	322	322
B	438	419
C	1773	1671
D	986	702
E	1468	1010

4.2 Independent Variables

We present here the independent variables included in our analysis. Diversity measures referred only to registered human users—to which we refer from now on as human users—because anonymous users often cannot be tracked across different edit sessions. Bot users were not included as well.

Tenure diversity. This variable was computed for each item by using the coefficient of variation [4] calculated on the number of days between each human user's first edit

[2] `https://www.wikidata.org/wiki/Wikidata:Item_quality`.

and the last day in our dataset. This method has previously been applied to measure tenure diversity in [2,8].

Interest diversity. The closeness of the editing patterns of users working on the same item has been used to estimate interest diversity, following the approach in [3]. To build this metric, we generated a two-dimensional matrix, in which all members of the group—all human users that performed at least one edit on the item considered—lie on one dimension and all items edited by anyone of them are on the other. Cells were assigned 0/1 values, according to whether a user had edited an item. The sparsity of the matrix—the ratio between the number zeros and the total number of cells—reflected how much the group members' editing patterns overlapped. The outcome values ranged from zero to one, with higher values indicating more diverse groups.

Proportion of bot edits. The proportion of edits made by bots over the total number of edits. This value was between 0 and 1.

Proportion of anonymous edits. The proportion of edits made by anonymous users over the total number of edits. This value was between 0 and 1.

Bot X Human edits. This variable was included to test how the interaction between bot and human editors affect outcome quality. It was computed by multiplying the proportion of human edits by the proportion of bot edits. Considering the low amount of anonymous contributions, this variable has values distributed in an inverted U shape, with higher values reflecting more balanced contributions from bots and humans.

Control Variables *Number of edits.* Items with a larger number of revisions are likely to have more statements and to have been reviewed and corrected more times.

Group size. The literature reports diverse effects of group size on outcome quality. Larger groups may negatively affect performance, because they reduce the likelihood of collaboration and increase the chance of conflicts [17]. On the other hand, more members likely entail a broader range of information [28]. We included group size as a control variable to account for these possible effects. Group size was measured by computing the number of unique editors for each item.

Age of the item. Older items have likely been seen, and reviewed, more times by editors. We used the number of days between the creation of an item and the last day in our dataset as a control variable.

4.3 Analysis Strategy

We performed an ordinal logistic regression (OLR) analysis to test the hypotheses. OLR takes into account the ordering of discrete response variables, such as the item quality labels used in this study, compared to other models which are either suitable to binary responses (standard logistic regression) or do not make any assumptions about the ordering of outcome discrete variables (multinomial logistic regression) [5]. OLR splits the distribution of the data corresponding to each rank in the response variable. It relies on the assumptions that independent variables have the same effect across different responses (*proportional odds* assumption) [6]. The ordering of these is modelled by considering cumulative probabilities for all different response categories, rather than by single category. As a consequence, the output of OLR provides an intercept value for each threshold between categories in the outcome variable.

We trained four models to predict item quality labels and verified the significance of the independent variables for prediction. The first model was the baseline and included only the control variables. Model 2 added variables related to the proportion of user type. Model 3 tested the influence of tenure and interest diversity, including only items that have ever been edited by humans in order to reduce sparsity of the data (tenure and diversity values were set to zero when no human users contributed to an item). Model 4 tested all the independent variables together, using all the items in our dataset.

4.4 Data

We accessed the complete dumps of Wikidata edit history[3], updated on 1st of April 2017, to provide data on every page in Wikidata at the time. We extracted from these dumps the completed revision history of each item in the labelled sample, including edit timestamp and user names. Only $4,987$ items over $5,000$ in the labelled sample were present in the dumps. Of these $4,124$ were edited by human editors. Scripts and data generated for the analysis have been made openly available online[4].

5 Results

Table 2 reports descriptive statistics of and correlations among the variables used in the analysis. The items in the sample greatly vary in terms of number of edits, group size, and item age. The ratio between edit number and group size shows that each user in a group carried out on average four revisions. If we consider the median item age (around four years) and number of edits, items are seldom edited. The proportion of registered human edits was, not surprisingly, highly correlated to bot edits, therefore it was left out from the models. Regarding diversity, items are edited by a population of editors which is moderately heterogeneous in terms of tenure. On the other hand, interest diversity was very high, indicating that on average editors focus on different sets of items.

The baseline model (1, Table 3) shows a positive significant influence of item age, edit number, and group size on item quality. The increase in quality level is very low for all three variables though, with item age having the smallest effect. Model 2 adds variables related to the contribution of different types of users to an item. The proportion of bot edits has a positive significant interaction with the response variable, thus **supporting hypothesis 1**. The influence of bots on item quality increases when these interact with human editors, as shown in Table 3, which **supports hypothesis 2**. The proportion of anonymous users is significant for prediction as well and influences negatively item quality. This means that **hypothesis 3 was supported**.

Model 3 was trained on items with at least one human edit. The distribution of quality labels for this set of items was more skewed towards higher levels, compared to the full dataset (Table 1). The results of model 3 show a significant positive interaction of tenure diversity with item quality (Table 4), thus **supporting hypothesis 4**. Interest diversity was as well a significant predictor, albeit with a lower positive influence on quality. Hence, **hypothesis 5 was supported**. Finally, model 4 included all

[3] `https://dumps.wikimedia.org/wikidatawiki/20170401/`.

[4] `https://github.com/Aliossandro/WD-group_diversity`.

Table 2. Descriptive statistics and correlations among independent variables. Item age is expressed in days since item creation.

	Mean	Median	Std	# Edits	*p* Bot edits	*p* Anonymous edits	*p* Human edits	Group size	Item age	Tenure div.
# Edits	135.4	28	239.19							
p Bot edits	.53	.50	.35	−.35						
p Anonymous edits	.01	0	0.03	.36	−.27					
p Human edits	.46	.50	.34	.32	−.99	.18				
Group size	36.32	7	57.48	.81	−.47	.49	.43			
Item age	1,182	1,507	557.16	.30	−.15	0.22	.13	.47		
Tenure diversity	.47	.38	.48	.40	−.49	.27	.48	.56	.62	
Interest diversity	.89	.98	.19	.11	.01	.12	−.02	.25	.16	−.12

the dependent variables. Significant interactions did not change, with the exception of the proportion of anonymous edits, which ceased to be a predictor of quality. The effect of group size decreases, compared with model 2. Moreover, tenure diversity had a stronger positive influence on quality, whereas the effect of the interaction between bots and humans decreases.

Table 3. Ordinal logistic regression of number of edits and group size, editor types, and diversity measures. Note: *** p< .001, ** p< .01.

	Model 1			**Model 2**		
	Coef.	SE	P	Coef.	SE	P
Label>= D	−.0715	.0609		−1.3024	.1037	***
Label>= C	−1.2553	.0642	***	−2.5499	.1081	***
Label>= B	–4.4452	.1028	***	–5.7677	.1361	***
Label>= A	–6.2173	.1320	***	–7.6024	.1628	***
Item age	.0003	.0001	***	.0001	.0001	
Group size	.0279	.0014	***	.0330	.0015	***
# Edits	.0029	.0003	***	.0033	.0003	***
p Bot edits				1.4005	.1029	***
Bot X Human				4.6909	.3377	***
p Anonymous edits				–3.8258	1.2218	**

6 Discussion

This paper has analysed the influence of group composition on outcome quality in Wikidata. First, we looked at how different proportions of bots, registered and anonymous human users affect quality. Second, we studied the effects of the distribution of two variables within groups of registered human users, tenure and members' interests.

The interaction between human editors and bots seems essential for the quality of Wikidata. It appears that the intertwinement of human and algorithmic contributions

Table 4. Ordinal logistic regression of number of edits and group size, editor types, and diversity measures, trained on items with at least one human edit. Note: *** p< .001, ** p< .01. Model 3 has been trained on the set of items with at least one registered human edit.

	Model 3			**Model 4**		
	Coef.	SE	P	Coef.	SE	P
Label>= D	−1.1739	.1779	***	−2.6487	.2125	***
Label>= C	−2.3874	.1815	***	−4.1062	.2175	***
Label>= B	−5.8900	.2145	***	−7.5732	.2450	***
Label>= A	−7.4843	.2262	***	−9.2759	.2573	***
Item age	.0002	.0001		−.0008	.0001	***
Group size	.0152	.0015	***	.0248	.0016	***
# Edits	.0039	.0003	***	.0040	.0003	***
p Bot edits				2.4695	.1237	***
Bot X Human				3.7688	.3618	***
p Anonymous edits				−3.6628	1.2403	
Tenure diversity	1.5502	.1104	***	2.8043	.1166	***
Interest diversity	1.0104	.1972	***	1.1004	.1999	

that led Niederer *et al.* to define Wikipedia as a socio-technical system [22] is also key for Wikidata quality. The division of work outlined in [21] may explain the strong positive effect of bot–human interaction on item quality. Each type of user contributes to Wikidata by carrying out the tasks in which they are specialised and require each other, in order to achieve good quality. Future work should investigate in detail this interaction at item level, focusing on which share of light– and heavy-weight tasks need to take on each user type, in order to successfully build an item. Fewer than half of the items in our datasets were ever edited by anonymous users. Although this reflects the overall edit distribution in Wikidata, this suggests that caution should be taken in interpreting results related to hypothesis 3 and that a more in-depth study should be conduct to draw clearer statements about that.

Heterogeneous groups in terms of tenure of their members are more likely to produce higher quality items. This contradicts prior studies around tenure diversity in an offline context, such as [2] and [23]. On the contrary, it agrees with the observations around Wikipedia in [16]. An explanation may be that in online contexts the importance of the relational aspect, which sees homogeneous groups perform better due to increased cohesion, decreases. More diverse groups would benefit from the different perspectives brought by their members, according to the information/decision making perspective [30]. This would apply specifically to Wikidata, where contrasting statements can coexist and editors do not need to discuss on talk pages to reach consensus, in contrast to Wikipedia, in which discussion pages are used to settle disputes. Another likely cause for the positive influence of tenure diversity on quality is the diversification of tasks carried out by users at different times of their activity within Wikidata [24]. The contributions of editors with various tenure levels may thus be complementary.

Our models show a significant interaction between interest diversity and quality. This finding is in agreement with previous research, which noted a linear correlation between this type of diversity and productivity [8] and between cognitive diversity and quality of decisions in Wikipedia [16]. Varied editor interests may imply that these are more active over the whole KG and know its mechanisms better. Furthermore, group editors that are active over a wider portion of Wikidata may have increased chances to link an item to others in the KG through statements. The interest diversity measure used does not take into account how conceptually distant the items edited by members of a group are. For instance, two users may engage in adding content related to British musicians, while still working on different items. Future work may rely on semantic similarity measures such as that presented in [25] in order to address this limitation.

Finally, the current study aims to shed light on the 'right mix' of users that leads to higher quality in Wikidata. According to the models trained, groups with higher levels of cooperation between bot and human editors (where tasks are more equally shared among these) are able to achieve better performance. 'Ideal' groups also benefit from including members with different tenure, which may address various quality issues. Group size has only a limited positive influence on performance, which partially contradicts previous observations around Wikipedia [15,16]. The presence of anonymous users in these groups seems marginal and does not have any significant effect.

7 Conclusions

This is the first research to address the relationship between group composition and outcome quality in Wikidata. Users of this system can be human, anonymous or registered, or bots. This investigation analysed how the contribution of these types of users and their interaction benefit Wikidata item quality. Furthermore, it examined the effects of tenure and interest diversity across registered human users on outcome quality. Ordinal logistic regression analysis revealed that the interaction between human and algorithmic users is necessary to create high quality items. Contributions from anonymous users are instead detrimental for quality. Concerning tenure and interest diversity, both these features have a positive influence on quality. More heterogeneous groups seem likely to benefit from the different experiences and skills of their members. One of the goals of the current study was to identify what are the characteristics of successful groups working on Wikidata items. These groups are slightly larger than average. Their members are both human and bots and contribute in a balanced proportion. Human editors in these groups are likely to have diverse levels of experience and interests in Wikidata.

Regarding the limitations of this work, cross-sectional approaches such as the one employed in the current paper may suffer from reverse causation and uncontrolled confounding factors [15]. Longitudinal analyses are effective for addressing these issues. Nevertheless, no measures of quality over time are currently available for Wikidata, to the best of our knowledge. This is a relevant research topic for the future of Wikidata and should be addressed by further studies. Several variables are at play in group work, such as the coordination among their members. Future research should explore how group diversity interact with other variables.

8 Acknowledgements

This project is supported by funding received from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 642795 (WDAqua ITN).

References

1. Adler, B.T., de Alfaro, L.: A content-driven reputation system for the Wikipedia. In: Williamson, C.L., Zurko, M.E., Patel-Schneider, P.F., Shenoy, P.J. (eds.) Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007. pp. 261–270. ACM (2007), http://doi.acm.org/10.1145/1242572.1242608
2. Ancona, D.G., Caldwell, D.F.: Demography and design: Predictors of new product team performance. Organization science 3(3), 321–341 (1992)
3. Arazy, O., Nov, O., Patterson, R., Yeo, L.: Information Quality in Wikipedia: The Effects of Group Composition and Task Conflict. Journal of Management Information Systems 27(4), 71–98 (2011)
4. Bedeian, A.G., Mossholder, K.W.: On the use of the coefficient of variation as a measure of diversity. Organizational Research Methods 3(3), 285–297 (2000)
5. Bender, R., Grouven, U.: Ordinal logistic regression in medical research. Journal of the Royal College of physicians of London 31(5), 546–551 (1997)
6. Brant, R.: Assessing proportionality in the proportional odds model for ordinal logistic regression. Biometrics pp. 1171–1178 (1990)
7. Brasileiro, F., Almeida, J.P.A., Carvalho, V.A., Guizzardi, G.: Applying a multi-level modeling theory to assess taxonomic hierarchies in wikidata. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 975–980. International World Wide Web Conferences Steering Committee (2016)
8. Chen, J., Ren, Y., Riedl, J.: The effects of diversity on group productivity and member withdrawal in online volunteer groups. In: Proceedings of the 28th international conference on Human factors in computing systems - CHI '10. p. 821. ACM Press, New York, New York, USA (apr 2010)
9. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandecic, D.: Introducing Wikidata to the Linked Data Web. In: The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I. Lecture Notes in Computer Science, vol. 8796, pp. 50–65. Springer (2014)
10. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Semantic Web (Preprint), 1–53 (2016)
11. Färber, M., Ell, B., Menne, C., Rettinger, A.: A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata and YAGO. Semantic Web 1, 1–5 (2015)
12. Harrison, D.A., Klein, K.J.: What's the difference? diversity constructs as separation, variety, or disparity in organizations. The Academy of Management Review 32(4), 1199–1228 (2007)
13. Haythornthwaite, C.: Crowds and communities: Light and heavyweight models of peer production. Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, HICSS (2009)
14. Jehn, K.A., Northcraft, G.B., Neale, M.A.: Why differences make a difference: A field study of diversity, conflict, and performance in workgroups. Administrative Science Quarterly 44(4), 741–763 (1999)
15. Kittur, A., Kraut, R.E.: Harnessing the wisdom of crowds in Wikipedia: Quality through Coordination. Proceedings of the ACM 2008 conference on Computer supported cooperative work - CSCW '08 p. 37 (2008)
16. Lam, S.K., Karim, J., Riedl, J.: The effects of group composition on decision quality in a social production community. Proceedings of the 16th ACM international conference on Supporting group work - GROUP '10 p. 55 (2010)

17. Levine, J.M., Moreland, R.L.: Progress in small group research. Annual review of psychology 41(1), 585–634 (1990)
18. Lukyanenko, R., Parsons, J., Wiersma, Y.F.: The IQ of the crowd: Understanding and improving information quality in structured user-generated content. Information Systems Research 25(4), 669–689 (2014), https://doi.org/10.1287/isre.2014.0537
19. Milliken, F.J., Martins, L.L.: Searching for common threads: Understanding the multiple effects of diversity in organizational groups. The Academy of Management Review 21(2), 402–433 (1996)
20. Moreland, R.L., Levine, J.M.: Socialization in organizations and work groups. Groups at Work: Theory and Research p. 69 (2014)
21. Müller-Birn, C., Karran, B., Lehmann, J., Luczak-Roesch, M.: Peer-production system or collaborative ontology development effort: what is Wikidata? In: OpenSym 2015 - Conference on Open Collaboration, San Francisco, US, 19 - 21 Aug 2015 (2015)
22. Niederer, S., van Dijck, J.: Wisdom of the crowd or technicity of content? wikipedia as a sociotechnical system. New Media & Society 12(8), 1368–1387 (2010), https://doi.org/10.1177/1461444810365297
23. Pelled, L.H., Eisenhardt, K.M., Xin, K.R.: Exploring the black box: An analysis of work group diversity, conflict, and performance. Administrative Science Quarterly 44(1), 1–28 (1999)
24. Piscopo, A., Phethean, C., Simperl, E.: 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) (2017)
25. Ribón, I.T., Vidal, M., Kämpgen, B., Sure-Vetter, Y.: GADES: A graph-based semantic similarity measure. In: SEMANTICS. pp. 101–104. ACM (2016)
26. Staab, S., Studer, R.: Handbook on ontologies. Springer Science & Business Media (2013)
27. Steiner, T.: Bots vs. wikipedians, anons vs. logged-ins. In: Proceedings of the companion publication of the 23rd international conference on World wide web companion. pp. 547–548. International World Wide Web Conferences Steering Committee (2014)
28. Surowiecki, J.: The wisdom of crowds. Anchor (2005)
29. Thakkar, H., Endris, K.M., Garica, J.M., Debattista, J., Lange, C., Auer, S.: Are linked datasets fit for open-domain question answering? a quality assessment. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics (WIMS16). ACM (2016)
30. Van Knippenberg, D., De Dreu, C.K., Homan, A.C.: Work group diversity and group performance: an integrative model and research agenda. Journal of applied psychology 89(6), 1008 (2004)
31. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57(10), 78–85 (2014)
32. Wagner, C.: Wiki: A technology for conversational knowledge management and group collaboration. The Communications of the Association for Information Systems 13(1), 58 (2004)
33. Yapinus, G., Sarabadani, A., Halfaker, A.: Wikidata item quality labels (5 2017), https://figshare.com/articles/Wikidata_item_quality_labels/5035796