Peer-production system or collaborative ontology engineering effort: What is Wikidata?

Claudia Müller-Birn
Freie Universität Berlin
clmb@inf.fu-berlin.de

Benjamin Karran
Freie Universität Berlin
benjamin.karran@fu-berlin.de

Janette Lehmann
Freie Universität Berlin
janette.lehmann@fu-berlin.de

Markus Luczak-Rösch
University of Southampton
m.luczak-rosch@soton.ac.uk

ABSTRACT

Wikidata promises to reduce factual inconsistencies across all Wikipedia language versions. It will enable dynamic data reuse and complex fact queries within the world's largest knowledge database. Studies of the existing participation patterns that emerge in Wikidata are only just beginning. What delineates most of the contributions in the system has not yet been investigated. Is it an inheritance from the Wikipedia peer-production system or the proximity of tasks in Wikidata that have been studied in collaborative ontology engineering? As a first step to answering this question, we performed a cluster analysis of participants' content editing activities. This allowed us to blend our results with typical roles found in peer-production and collaborative ontology engineering projects. Our results suggest very specialised contributions from a majority of users. Only a minority, which is the most active group, participate all over the project. These users are particularly responsible for developing the conceptual knowledge of Wikidata. We show the alignment of existing algorithmic participation patterns with these human patterns of participation. In summary, our results suggest that Wikidata rather supports peer-production activities caused by its current focus on data collection. We hope that our study informs future analyses and developments and, as a result, allows us to build better tools to support contributors in peer-production-based ontology engineering.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: Computer supported cooperative work—web-based interaction

General Terms

Wikidata, peer-production, collaborative ontology engineering, participation patterns

OpenSym '15, August 19 - 21, 2015, San Francisco, CA, USA.
ACM 978-1-4503-3666-6/15/08 ...$15.00.
http://dx.doi.org/10.1145/2788993.2789836.

1. INTRODUCTION

The Wikidata project aims to create a free, structured knowledge base that can be read and edited by humans and machines alike. Wikidata has its origin in Wikipedia, the world's largest peer-produced encyclopedia, with the particular purpose to manage facts represented in Wikipedia articles. Such a fact is, for example, the number of inhabitants of the *United States*. The challenge is that this fact exists not only in the article of the English language version, but also in 264 other languages. In English Wikipedia, the United States have 320,925,485 inhabitants; however, in German Wikipedia, it has 317,238,626 and in Spanish Wikipedia, 316,017,000 inhabitants.

Which is the correct number? The Wikidata community addresses such factual inconsistencies by providing one central place where such data is stored. Instead of having the population of the United States maintained in each language version separately, Wikidata holds the number and will provide it automatically to all Wikipedia language versions. Thus, the number of inhabitants only has to be stored and updated in Wikidata.

The need for a structured knowledge base such as Wikidata has emerged over the course of the last decade. During that time, the role of structured data in Wikipedia has changed. At first, the Wikipedia community itself started to provide some structured data manually. Editors made use of standardised infobox templates in specific articles, such as biographies. These present typical information that most if not all articles in this area should contain in a tabular format.

Later, the DBpedia[1] project began exploiting this feature by automatically extracting structured data from Wikipedia infoboxes and building a self-contained structured knowledge base. The extraction algorithms are configured via a dedicated wiki system that allows to define mappings between Wikipedia infoboxes and DBpedia's terminology. In contrast to the terminology, which is maintained by the team of academics behind DBpedia, the mapping wiki is open to volunteers' contributions.

An alternative approach pursues projects such as ICD-11[2]. The ICD-11 has been created to allow users to share and compare data about diseases in a consistent way. The project aims

[1] http://dbpedia.org

[2] The ICD-11 is the International Classification of Diseases of the World Health Organization (WHO).

at creating a knowledge base collaboratively that is grounded in conformity with a specific terminology. Such terminology is defined by conceptual knowledge which consists of classes (or sets) with their properties and relationships [13]. A "disease", for example, can have a "temporality" property. The structured data is represented in concrete instances of these classes. The property values (e.g. possible values for "temporality" are acute or chronic) of a "disease" class are instances of the class "Term" [29].

Wikidata appears to sit between the manual approach of Wikipedia's peer-production community, the automatic knowledge extraction approach of DBpedia and the collaborative ontology engineering approach of the ICD-11. In our research, we assume that Wikidata combines the characteristics of a peer-production system and a collaborative ontology engineering project. Currently, research in both areas tends to be separate.

Our knowledge of Wikidata and the existing synergies possible between manual and automatic processes of content production is still somewhat limited; modes of production and viable implications for similar systems are not yet known. We combine both perspectives with our research and make the following contributions:

- We introduce a scheme of categorizing edits into action sets on Wikidata.
- We use this scheme to identify participation patterns by means of clustering for human and algorithmic contributions.
- We discuss these contribution patterns in light of existing roles in peer-production communities and ontology engineering projects.
- We examine different lines of development for which Wikidata might strive in the future.

The remainder of this paper is structured as follows: In the next section, we review existing literature. In Section 3, we introduce the Wikidata project and explain our data. Following in Section 4, we describe the process of identifying actions and resulting action sets on Wikidata. In Section 5, we use these to show the structure of human and algorithmic participation on Wikidata. In Section 6, we discuss our results and link them to existing research. Finally, we examine the limitations of our study and highlight future research directions in Section 7. We conclude with a more general perspective on this research.

2. RELATED WORK

We can describe the Wikidata community from two different perspectives: (1) the peer-production community perspective, and (2) the collaborative ontology engineering perspective. The first of these perspectives tends towards the creation and maintenance of a valuable artefact [7]. In this regard, scholars have mainly studied peer-production communities in contexts ranging from open source software development (Apache as an early analysed example, *e.g.* [20]) to textual knowledge bases (including Wikipedia as the most prominent example, *e.g.* [23]). The latter perspective has mainly been studied by scholars in Knowledge Engineering and the Semantic Web, who focus on ontology engineering methods and tools (*e.g.* [16, 28]). We hypothesise that Wikidata *combines* both perspectives and, as such, the following sections give a brief overview of both lines of research.

2.1 Peer-production communities

Participation in peer-production communities is not evenly distributed – people often start with a peripheral level of participation. As their experience grows, it appears that their contribution will often rise in tandem [17, 32]. This development is often described as a layered model. The different layers represent distinct roles in the community and depict how people are positioned, allowing some conclusions to be drawn from contributors' activities [25]. Ye and Kishida [33] describe the travelling through these different layers as "role transformation". They define different roles that may be represented in open source software development projects, starting at the outermost circle with the passive contributor and ending in the innermost circle with the project lead [22, 33].

Liu and Ram [18] identified various types of actions editors can carry out in Wikipedia and used these actions to classify contributors into groups. In another piece of research, roles have been defined based on the page editing activities of contributors in different namespaces[3] [31]. A matching of roles to a contributor's activities reflects the principle that exists in many peer-production communities: A contributor can only earn a role by merit. Preece and Shneiderman [24] generalise this line of research into the *Reader-to-Leader Framework*. Both the amount and type of participation appear to change over time. At the beginning of the Wikipedia project, for example, most of the editing work was carried out in the main namespace. Over the course of the project, focus shifted to other namespaces, the community namespace being one example [14].

The same can be seen regarding people's participation. Bryant *et al.* [5] finds that contributors change their type and scope of participation with increasing community involvement. Novice editors tend to concentrate on a single article or a particular set of articles, whereas expert contributors expand their activities to the health of the overall project and even support the community by adopting a number of different roles [5].

Roles allow the description of the participatory architecture of a peer-production community. When community members change the regularity and type of their participation, they also change the social dynamics, and reshape the structure of the community [22, 33]. A balanced composition of all roles is important to ensure its sustainable development [20]. An improved understanding of these roles can help one to understand how a peer-production community co-ordinates its collective actions, and provides a gauge to measure the current and future health of that community [31].

2.2 Collaborative ontology engineering

Over the last few years, collaborative ontology engineering that allows distributed and disparately skilled teams to build ontologies has gained increasing momentum and led to a new family of ontology engineering tools (*e.g.* Semantic Media Wiki [16], Collaborative Protegé [28]). A recent study shows that two abstract roles are commonly represented across the collaborative approaches [26]: (1) ontology editors who are able to change the ontology directly, a task formerly assigned solely to ontology engineers, and (2) ontology contributors who are restricted to providing feedback or suggesting changes.

[3] All pages in Wikipedia are organized into namespaces. The main namespace, for example, contains the encyclopedic articles.

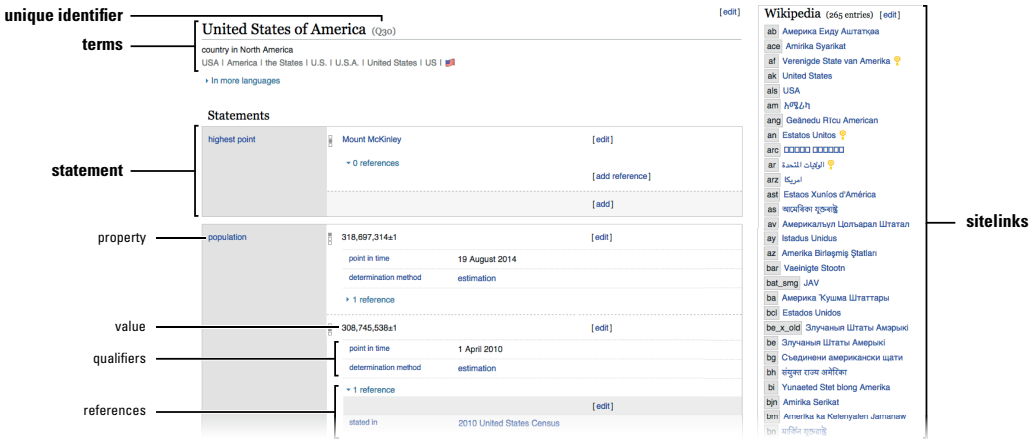

Figure 1: Wikidata's user interface showing the item Q30.

In contrast to traditional ontology engineering, collaborative approaches have changed the way in which the conceptual knowledge layer (*i.e.* classes with properties and their relationships) is presented to the contributors. Contributors can, for example, discuss the creation of new classes in Collaborative Protegé. Research often focus on projects in which people with a dedicated education and organisational involvement are allowed to make changes to knowledge bases. Although these knowledge bases are large (*e.g.* the ICD-11 ontology consists of over 33,000 classes), the number of people involved in the development is small (ranging from 5 to 76 contributors [27]). In addition, the process that leads to the creation of these knowledge bases is admittedly collaborative, but not everybody can participate.

Falconer *et al.*, for example, analysed the development processes of three large-scale biomedical ontologies using Collaborative Protégé [28]. Although there is an increased flexibility in being able to switch between roles, practice shows that there is still a fine-grained differentiation of roles at a task level [9]. Every participant is an active ontology editor, but whether they created new classes, properties or instances (*i.e.* closeness to semantic concepts) differed greatly.

Other research focuses on Semantic Wikis, such as the Semantic MediaWiki [16]. Although this software is open to everybody and has a "human-readable interface[...] to ontologies" [4], the knowledge base is often limited to one specific topic domain. Gil *et al.* [12] analysed editing actions in various Semantic MediaWiki communities and differentiated roles depending on whether they contributed to the ontology (*e.g.* to classes or properties) or not (*i.e.* contributors provide instances). Their results suggest that building structured data based on Wiki-software is a very individual endeavour and that a common approach does not exist – each community develops its own best practice on how to use conceptual knowledge in their specifications.

Freebase[4] is probably the most prominent example of an open system in which contributors can build a structured dataset to describe topics of interest and create collections of interlinked topics [3]. However, existing patterns of participation and how people dealt with ontological primitives will remain unknown. In December 2014, Google announced the termination of this service by June 2015 and all data has now been transferred to Wikidata. Henceforth, Wikidata is the only existing example of an open community that allows the creation of structured data by anybody. We next introduce this project in more detail.

[4]http://www.freebase.com/

3. WIKIDATA PROJECT

We have already introduced Wikidata as a structured knowledge base, that aims to curate consistent and high-quality data across the various Wikipedia language versions. The Wikidata approach is based on a number of design decisions, such as open editing, plurality (conflicting data being allowed) and multilingual data [30]. These principles differentiate Wikidata from other collaborative ontology engineering projects. The ICD-11, for example, does not allow public editing, Freebase has no multilanguage support for its entities and Semantic MediaWiki makes it challenging to capture references to external sources [30].

The main concepts of Wikidata can best be explained by the content model[5] employed and its representation on the user interface. Both are introduced in the next section. Subsequently, we describe the datasets used in this study and present a number of content and community-related characteristics.

3.1 Content model

Wikidata aims to provide a clear user interface that gives various opportunities to participate easily in collecting structured data without having previous knowledge of such data. Similar to Wikipedia, Wikidata is organised in pages. Each page corresponds to an item, situated in the item namespace, or a property, organised in the property namespace. We focus our study on these two namespaces, as they represent the result of the community effort – the structured dataset. These structured data consist of conceptual and instance knowledge. The conceptual knowledge is represented by classes of items (*e.g.* geographical object, mood) and properties (*e.g.* highest point, instance of). The instance knowledge is represented by concrete items, such as "United States of America" or "happiness". Both items and properties have a unique identifier (*e.g.* Q30 and P17, respectively) and are described by terms and statements.

[5]A detailed description of Wikidata's content model is given by Erxleben *et al.* [8].

Item namespace		
# Items		16,503,623
Terms	# Labels	58,500,696
	# Descriptions	41,040,885
	# Aliases	4,707,583
# Statements		46,326,190
	# Claims	49,785,283
	# References	26,271,077
	# Qualifiers	817,879
Property namespace		
# Properties		1,281
Terms	# Labels	38,013
	# Descriptions	16,525
	# Aliases	4,828

Table 1: Content-related statistics of Wikidata of the item namespace and the property namespace.

	Item namespace	**Property namespace**	**Property talk namespace**
#contributors	296,367	2,358	525
# anon. users	227,724	964	46
# reg. users	68,481	1,390	474
. . . # active users	20,163	643	158
# bots	162	4	5
# revisions	164,086,942	76,497	14,921
of anon. users	760,029	1,945	84
of reg. users	24,007,629	74,423	14,169
. . . of active users	23,778,186	69,316	12,330
bots	139,319,284	129	668

Table 2: Community-related statistics per namespace with the number of contributors for each group and their revisions.

In addition, items can have site links which connect them to pages on other Wikimedia sites, mainly to various different language versions of Wikipedia, but also to other Wikimedia projects. Figure 1 shows Wikidata's user interface and content elements for items. We will go on to describe each element in more detail. Since properties exhibit the same structure, they are not described further here.

Terms are language-specific and refer to labels (*e.g.* United States of America), descriptions (*e.g.* country in North America) and optional aliases (*e.g.* USA). They are mainly used for displaying items on Wikidata, depending on the language setting of the user. Statements describe items by their characteristics and in their simplest form consisting of a property and at least one value (*e.g.* a string, time, co-ordinates or another item). As shown in Figure 1, the property "highest point" has the value "Mount McKinley" which refers to another item. This property-value pair is called a claim, and a statement can consists of various claims. The statement of the property "population", for example, has 27 claims. This feature is especially important since it allows even conflicting values to be stored. The rank of the claim provides a way to indicate the preferred value. It is shown by a small rectangle on the left side of the claim (cp. Figure 1). More complex claims are made possible by adding references, to describe the origin of the data, and qualifiers, to provide contextual information.

3.2 Datasets

Our study is based on two Wikidata dumps generated on November 3, 2014. The first dump contains all Wikidata pages without their history serialised as JSON.[6] The second dump file is an XML serialisation of the full history of all pages in the Wikidata wiki.[7] Since the project officially launched on October 29, 2012, we trimmed our dataset from this data to October 29, 2014.

Content characteristics. An overview of the characteristics of our dataset is given in Table 1. Nearly 17 million items exist that are described by over 60 million labels showing Wikidata's multilingualism. These items are described by over 45 million statements indicating that, on average, each item is described by at least two statements. However, the number of references is considerably low and they refer mainly to Wikipedia.

Community characteristics. Wikidata's open editing approach makes its community very similar to Wikipedia. We can differentiate three contributor groups: (1) anonymous users (*i.e.* not registered or logged-in users), (2) registered users (include active users, *i.e.* registered users with at least five edits per month) and (3) bots.[8]

We identified the bots in our dataset as of March 10, 2015, as follows: We extracted all users labelled as "bots" in the `user_group` table from the Wikimedia Tools Labs.[9] We also examined manually a second list of users that are considered to be bots but are not listed in the `user_group` table.[10] This resulted in 162 bots.

In addition to the item and the property namespace, we considered the property talk namespace for the following reason. An important activity in collaborative ontology engineering is the users' discourse about conceptualisations, for example, properties in Wikidata. In the latter, this takes place on talk pages in the property space[11]. These talk pages also contain constraints with regard to their usage. We refer to the property talk namespace as the discussion namespace.

Table 2 reports various statistics for each contributor group for all three namespaces. About 230,000 anonymous users (identified by distinct IP addresses), 70,000 registered users, and 160 bots carried out nearly 165 million edits on Wikidata in our analysis period.

By looking at the item namespace, we could ascertain that about 30 percent of the registered users had made more than five edits in each month. These user group created most of the revisions that were made by humans. We see, however, that the vast majority of edits in the item namespace are carried out by bots, which perform 85 percent of all revisions.

Further analyses show that the number of edits per contributor (bots and users) is highly skewed. It means only few

[6] https://archive.org/details/wikidata-json-20141103

[7] http://dumps.wikimedia.org/wikidatawiki/20141106/

[8] Bots are software programmes that perform edits autonomously following their own predefined schedule.

[9] http://tools.wmflabs.org/

[10] https://www.wikidata.org/wiki/Category:Bots_without_botflag

[11] Only 0.02 percent of item pages have a corresponding talk page.

	# $Edit_{Item}$	# $Edit_{Prop}$
create (item or property)		
create/merge/clear/redirect	15,514,688	1,276
set statement		
add/change/remove claim	58,825,720	–
add/change/remove reference	15,229,192	–
add source	13,903,977	–
add/change qualifier	95,792	–
change rank	1,201	–
set term (of item or property)		
add/change/remove label	39,818,995	45,096
add/change/remove description	13,650,353	21,214
change/remove alias	399,572	7,764
set sitelink		
add/change/remove sitelink	6,001,811	–
move Wikimedia site	438,396	–
change badge	84,079	–
protect (item or property)		
protect	232	19
revert (item or property)		
revert	121,236	1,128

Table 3: Action sets with underlying actions, differentiated into item-related and property-related action sets.

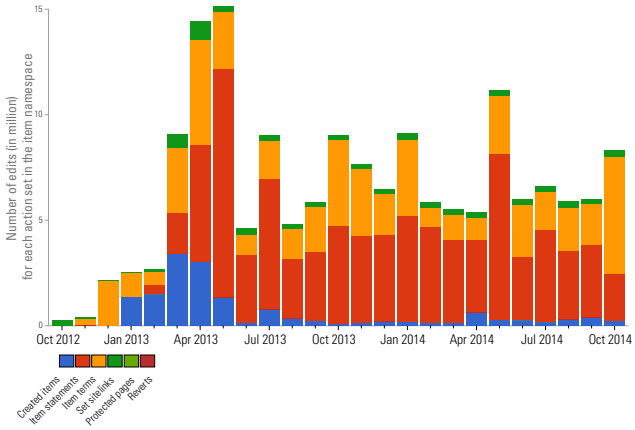

Figure 2: Number of edits (in millions) in each action set in the item namespace over the period of analysis. The number of reverts and protects are small, and thus, barely visible.

contributors are responsible for the majority of edits.

Moreover, only a minority of users participate in the creation and maintenance of properties, yet over 45 percent of these are active in this namespace on a regular basis. A similar pattern is visible in the property talk namespace. In both namespaces, only a minority of edits is carried out by bots. The low number of bots in property-related activities might suggest a division of labour within the community. In the next section, we look more closely at the number and type of edits in Wikidata.

4. TYPES OF PARTICIPATION

Our goal is to shed light on the types of contributions on Wikidata, allowing us to sketch a more detailed picture of the participation. We build on the research carried out in the context of Wikipedia [15, 18] by looking at two types of data that reflect collaborative activities in Wikidata: (1) edit histories provided for each change a contributor performs (who, when), and (2) edit information that is automatically created when a contributor performs a change. Following that, we describe the basis for categorising contributions on Wikidata.

4.1 Identifying actions on Wikidata

We analysed a random sample of the revisions of items and properties on Wikidata and discovered that edits contain detailed contextual information about the amended data. We took advantage of this information in order to differentiate edits more precisely. Each edit in Wikidata has a comment that describes the particular action carried out on a particular scope of an item or a property, and the actual data that was added. We applied several steps to identify the types of edits that are described in the following:

(1) Normalisation: Firstly, we parsed all edits in our dataset and normalised their comments by removing all digits and everything after `*/`. This essentially left us with the type of an edit, ignoring information about what was edited. Additionally, we checked if comments started with certain known phrases, such as `rv` or `undid`, and ignored everything that followed. Furthermore, we aggregated the shortened comments. The comment `/* wbsetlabel-set:1|en*/ Africa`, for example, was normalised to `/* wbsetlabel-set`.

(2) Labelling: We then looked up every identified comment field manually in Wikidata and translated the system's code into a human readable version. `/* wbsetlabel-set`, for example, was labelled "Changed LANGUAGE label".

(3) Coding: In order to categorise all comments into *actions*, two members of our research team aligned all the codes into consistent verb-noun pairs. After they had coded a subset of comments individually, they discussed ambiguous cases and adjusted actions until a consensus was reached. The label "Changed LANGUAGE label", for example, was assigned the action "change label". Based on this classification, we observed that actions form groups based on their scope. As such, it is possible to differentiate between actions that are related to statements alone, and others related to information represented in the header section of an item (*i.e.* to terms).

(4) Aggregation: We decided to aggregate all codes that belonged to the same kind of action into just one *action set*. As an example, every change in term (label, descriptions, or aliases) of an item was summarised into *set term*. The previous example action "change label" belongs to the action set *set term*, too.

This process resulted in six distinct action sets that are listed in Table 3, which also shows the actions after the first round of coding in the rows below.[12] Based on this process, we classified over 99 percent of all edits in the item namespace (1,698 edits remain unclassified) and all edits in the property namespace in our dataset.

[12]Adding statements to properties was enabled by November 2014. Our period of analysis ends in October 2014. As a result, our analysis did not capture this community activity.

4.2 Evolution of the edit count of action sets

The edit count for the action sets varied over the course of the project. Figure 2 depicts the evolution of the edit count of action sets in the item namespace over the period of analysis. At the beginning of the project (five months in duration), the edits were comparatively low, since only selected Wikipedia language versions (*e.g.* Hungarian Wikipedia) used Wikidata for testing. In April 2013, the largest language version of Wikipedia – English – started using Wikidata language links. During this time, contributors focused on representing Wikipedia articles as items in Wikidata, describing them with terms and linking them to the respective language version. From May 2014, activities relating to statement editing represented over 50 percent of all edits in Wikidata. The question is whether contributors contributed primarily to either one or multiple action sets or not. To begin answering this question, we used the action sets for identifying typical participation patterns on Wikidata. The results of our analysis are described in the following.

5. STRUCTURE OF PARTICIPATION

We use our classified edits from Section 4.1 to determine different forms of participation in Wikidata. Firstly, we describe the procedure employed to analyse the data, and secondly, we present our results.

5.1 Study procedure

We carried out the analysis of participation patterns in four steps:

(1) Data aggregation: We transformed our data into monthly time frames per contributor (*contributor time frames*), meaning essentially that all contributor edits within a one month time interval were aggregated. This resulted in 222,093 contributor time frames from 68,704 contributors, who carried out 164,178,360 edits.

(2) Contributor selection: We then discarded all time frames of contributors who made fewer than five edits to analyse only the active contributors. Since the tracking of anonymous users is hard, we removed them entirely from our dataset. This culminated in 89,814 contributor time frames, which still represented 99.4 percent of all edits within the time period analysed.

(3) Feature computation: We used the action sets described in Table 3 to represent contributor activity on Wikidata. For each contributor time frame, we counted how many actions a contributor performed in every action set in the item and property namespace. We learnt in Section 3.2 that the number of edits per contributor is highly skewed. Thus, instead of taking the absolute value, we normalised the values for each contributor time frame by the total number of actions for that contributor time frame and namespace. We determined, for example, how often a contributor created an item in a month and divided this by the number of all edits made by that contributor in that respective month in the item namespace. We joined reverts and page protections across namespaces, since those tasks have the same meaning across all namespaces (cp. Table 3). Moreover, we extended the feature set with the feature *property discussion* (cp. Section 3.2). We computed the relative value for this feature again, as described above.

These considerations resulted in the following feature set, computed per contributor time frame: *created items*, *item terms*, *item statements*, *set sitelinks*, *created properties*, *property terms*, *discussions*, *protected pages*, and *reverts*.

(4) Pattern detection: Based on the features selected, we clustered the human and bot time frames using the k-means algorithm. We applied the algorithm repeatedly and incremented the number of clusters in each iteration. To determine the number of clusters to continue with, we computed the stability of each clustering round according to Ben-Hur *et al.* [2]. In the latter work, stability is determined by performing several clusterings using random subsamples of data, followed by an analysis of the distances between the clusterings. We performed the clustering nine times on random subsamples of 80 percent of the data. We employed the normalised variation of information measure (*VI*) [19], whereby $VI = 0$ indicates that all objects are in the same cluster (full stability) and $VI = 1$ indicates that all objects are in different clusters (no stability), as a distance measure.

We chose six clusters for humans, as we observed a notable increase in stability with six clusters that did not improve much for seven clusters. The average normalized VI for six and seven clusters is 0.07 ($sd = 0.04$) and 0.05 ($sd = 0.05$), respectively. We selected the same number of clusters for the bot time frames clustering, as it allowed us to compare clusterings at the same level of granularity. Moreover, we observed a decrease in stability from six to seven clusters, respectively. This showed that our clustering was stable.

5.2 Patterns of participation

In this section, we introduce each of the participation patterns identified for both humans (*i.e.* users) and bots. We deliver a more in-depth scrutiny of the patterns in the discussion section.

Our analysis described previously resulted in six clusters for humans and five clusters for bots, which represent typical patterns of participation. The sixth bot cluster could not be interpreted as a typical participation pattern, because it included only one time frame and, as such, we excluded it from our result presentation and discussion.[13] The remaining five bot patterns are very similar to the human patterns. It is only the Property Engineer that is unrepresented in the bot patterns.

Figure 3 shows the participation patterns where each column refers to one specific pattern indicated by name. Besides the cluster center (normalized by the z-score), additional information is provided that gives further insights into each pattern. Firstly, we show the absolute and relative numbers of contributor time frames for each pattern that correspond to the size of the cluster. Secondly, the number of contributors following each pattern is represented. It is worth mentioning once again that our data is based on contributor time frames and, as such, contributors can belong to more than one pattern. Thirdly, we show the absolute number of edits for each pattern (total edit count) as well as the average number of edits a contributor has carried out per time frame (mean edit count). Finally, the contributor activity distribution diagram is included, where each bar represents the relative number of contributors showing the respective participation pattern in a three-month period.

[13]The time frame represents the JWbot which reverted a high number of incorrect edits in the time frame 2014-08.

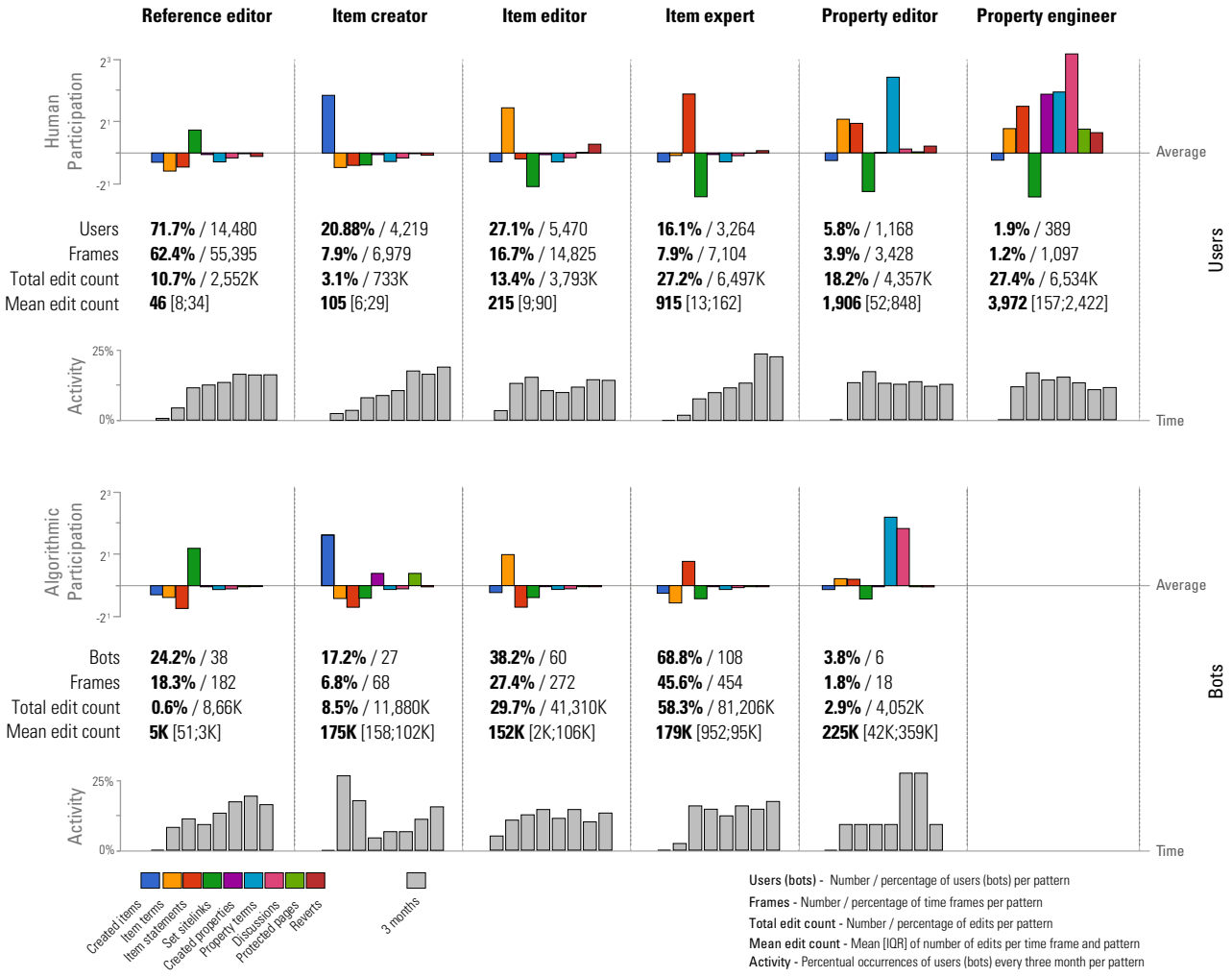

Figure 3: Participation patterns on Wikidata for humans (top) and bots (below), showing for each pattern the cluster centers, basic statistics (*i.e.* relative number of contributors and time frames as well as total and mean edit count) and the activity distribution diagram.

Looking at the general properties of all participation patterns, we observed that bots have several magnitudes more edits on average than humans. Additionally, the distribution of edits per contributor is heavily right skewed for most patterns, as indicated by the mean falling outside to the right of the interquartile range (IQR). The only exception is the Property Editor bot pattern. This indicates that a few extreme contributors form the participation pattern.

Next, we briefly introduce these participation patterns for humans and bots together and mention differences where necessary.

Reference Editor: The primary activity of a Reference Editor in Wikidata is to add sitelinks to Wikipedia and other Wikimedia projects. On the one hand, the majority of humans in our dataset carried out this activity during their time of participation. On the other hand, only a quarter of the bots slipped into this pattern. For both, the number of contributors increased slightly over time, but the total and mean edit count are very low. This indicates that humans and bots carry out this activity often, but only selectively.

Item Creator: Contributors that follow this pattern are concerned primarily with the creation of items, whereas all other actions are roughly average. In contrast to humans, bots have a slightly above average value for created properties and protected pages. The Item Creator represents the third largest pattern in terms of humans included and the fourth largest for bots. For humans and bots, the number of contributor time frames is comparatively low, indicating more sporadic participation in activity. The mean edit count is very low for humans, whereas for bots it is mid-range. Over time, the number of humans that belong to this pattern increases, as can be seen in the activity distribution diagram. Bots show a similar pattern, with the exception of a peak at the beginning of our analysis period, probably caused by Wikidata's inception phase.

Item Editor: Contributors who primarily edit terms from items are summarised in this pattern. They are Item Editors and probably carry out these edits in their respective user language. Bots that follow this pattern appear to support humans in creating terms on items at a constant level. Other

actions, especially *set sitelinks*, are well below average. The mean edit count for bots is the second highest. The IQR of the mean edit count shows that a majority of humans have a low editing activity. For humans and bots alike, the number of contributors that belong to this pattern is stable over the period of analysis.

Item Expert: The main action of an Item Expert is to *set statements on items*. This corresponds to the task of adding the conceptual context to items within a domain of interest, such as typical properties of items (*e.g.* the "country" "United States of America" has a "capital"). The action *set sitelinks* is notably below average. The number of contributors and time frames is low for humans compared to the other patterns, but maximum for bots. The total edit count is higher and the mean edit count is mid-range for humans and bots. For humans, however, the activity distribution diagram shows an increased number of contributors, indicating that more of them are involved in the activity. On the contrary, this value remained approximately constant for bots.

Property Editor: The Property Editor pattern represents contributors who are mainly involved in setting terms on properties. Their activities in the item namespace concerning terms and statements are also above average. In addition to humans, bots show well above average activity in discussions. For humans, this is the second smallest pattern and the smallest for bots. For both, the number of contributors is almost evenly distributed over time and the mean edit count is very high. Bots, however, show an activity peak to the end of the analysis period.

Property Engineer: Contributors in the Property Engineer pattern are especially active in the property namespace. They are mainly involved in the creation of new properties (conceptual knowledge) and also in describing properties and editing talk pages. The result of these property creation processes is reflected in the user's participation in defining statements and also terms on items. This pattern is the only one that exists for humans and has no corresponding bot pattern. Though only a minority of humans correspond to the pattern of Property Engineer, the mean edit count is very high as well as the total edit count. In addition, the number of humans is distributed almost evenly over time.

We discuss these results in more detail in the following and summarize them into three main insights.

6. DISCUSSION

The starting point in identifying participation patterns on Wikidata has been a set of classified edits (cp. Table 3) derived from the edit history comprising 99 percent of all edits within the period of analysis (October 2012 to October 2014). These actions allowed us to describe the behaviour of human and algorithmic contributors in Wikidata, in order to identify overlapping and varying areas of activity. We identify six mutually exclusive participation patterns that best describe our underlying dataset. Our results evolve around three themes that are discussed next.

6.1 Specialised or generalised contributions

The majority of the resulting patterns (four out of six) consist of one specific action set that acts as a decisive factor for pattern creation. This applies especially to the Item Creator and the Reference Editor pattern. Both show only one outstanding feature, whereas all other feature values are almost average. The other two patterns (Item Expert and Item Editor) show deviant behaviour in reference additions on items. People editing items (*e.g.* adding descriptions or statements) rarely add links to other Wikimedia projects. These results suggest that editors seem to be very task-focused over at least a one-month time period. This specialised contribution behaviour might be supported by Wikidata's structured form-based user interface. However, Liu and Ram [18] show similar results in Wikipedia. They align their analysis to Wikipedia's content model and, as a result, they present contributors' content contributions similar to ours. It seems to be a typical division of labour for peer-production systems.

Two patterns show a very distinct contributing behaviour: the Property Editor and the Property Engineer. In both cases, people carry out edits in the property namespace as well as the item namespace. The average number of edits per contributor in these participation patterns is the highest, but the IQR shows very unequally distributed edits. This unequal distribution looks to be caused particularly by contributors who have special software rights, namely those who are able to create new properties on Wikidata[14]. 14 users belonging to the Property Engineer pattern have this special access right which corresponds to all users with that right. According to Arazy *et al.* [1], functional roles such as this can be seen as a career path in a peer-production community. As opposed to administrators, who are "responsible for the social administration of the [...] community", contributors that are allowed to create properties can be seen as content administrators. Those contributors especially belonging to the Property Engineer pattern have probably the best understanding of semantic concepts (*i.e.* classes with properties and their relations) within the community, notwithstanding their being the smallest participation pattern in terms of contributor numbers.

6.2 Human vs. algorithmic contributions

A special characteristic of Wikidata is that all participation patterns identified have an algorithmic counterpart. The participation patterns of bots align with the human patterns. A major difference is the total amount of bot activities, which differs considerably from the human activities, which is expected since bot activities scale better.

Even though users belonging to the Item Expert pattern account for one third of all edits, bots carried out almost 60 percent of all edits in this area. These bots are often responsible for linking to authorities' ids (*e.g.* VIAF) and identifiers (*e.g.* MusicBrainz artist ID), which resulted in a high number of edits. The ProteinBoxBot, for example, adds claims and their references based on gene information from existing online databases.

This indicates a division of labour on Wikidata. Geiger and Halfaker describe similar findings in their research on Wikipedia. They show how a distributed cognitive network of human and algorithmic actors work efficiently together to detect and revert vandalism [11]. Compared to Wikipedia, which has had several years to develop a sophisticated editing system (*e.g.* also in terms of its governance [6]), the Wikidata community is very young. Our results indicate that bots are particularly responsible for creating items in the first six months and for subsequently editing statements.

We assume that these types of human-algorithmic co-op-

[14] http://www.wikidata.org/wiki/Wikidata:Property_creators

eration might change as project maturity increases. At the moment, the human and algorithmic contributors seem to focus primarily on adding content.

One might argue that the involvement of algorithmic contributions might decrease when all links to external knowledge bases are added. However, this is very unlikely. We assume the community will develop a more sophisticated system in which algorithmic contributors are more deeply involved in the project maintenance, as has already been shown in the Wikipedia context [21]. The question is, to what extent bots will be involved in the more sophisticated knowledge building process (*e.g.* checking of logical inconsistencies in the conceptual knowledge layer).

6.3 Peer vs. ontology production

Our results give initial insights into the social architecture of the Wikidata community regarding existing participation patterns. Even though our results are not conclusive, they show that Wikidata finds itself between two approaches – "classic" peer-production and collaborative ontology engineering.

Taking the ontology engineering perspective, we observe that Wikidata softens the boundary between the conceptual and instance level. The majority of contributors work on items that can either be instances (*e.g.* United States of America (Q30)) or classes (*e.g.* country (Q6256)). This implies that the semantic closeness of the contributions cannot be differentiated as easily as in other ontology engineering projects. In Freebase, for example, an explicit distinction is made between instances and classes.

Another approach to analyse the different kind of contributors in Wikidata is to evaluate the previous knowledge required to carry out specific edits. In the case of the Reference Editor and the Item Editor, for example, little or no semantic knowledge is needed to describe or translate and link items primarily to Wikipedia. The user *Condor3d*, for example, is a typical member of the Reference Editor pattern which simply links items to the Dutch Wiki. This differs from other contributors, such as the Item Expert. Here, contributors need to understand the structure of statements in order to make useful contributions. It seems to be even more challenging for inexperienced contributors to contribute to the property namespace, indicated by the low number of contributors in the Property Editor pattern and the Property Engineer pattern.

Indeed, Wikidata is challenging on modelling the conceptual level. Contributors mention in a recent discussion on the Mailing list that they use Collaborative Protegé (cp. [28]) for modelling classes and properties instead of designing their data structures directly on Wikidata. It seems that the simplified user interface of Wikidata is very valuable for data contributions, but less valuable for data modelling. It hinders contributions from the community that focus on the conceptual level rather than the instance level. At the moment, peer-production aspects are mainly supported, probably caused by Wikidata's origin. The years to come will probably show in which direction it is being developed. Wikidata's community is evolving, however, and new decisions that can also influence contribution patterns emerge every week.

The identification of the aforementioned participation patterns allows us to increase our understanding of the process of creating structured data on Wikidata. We hope that our study informs future analyses and developments and, as a result, allows us to build better tools to support different contributor behaviour in peer-production-based ontology engineering.

7. LIMITATIONS AND FUTURE WORK

Our study has several limitations worth mentioning. Firstly, by defining action sets, we regard every edit as being equal. We account neither for the persistence of an edit nor for the type of action (add, change or remove). Since our main interest was in locating types of participation on Wikidata, we aggregated this information. An alternative approach for counting edits on Wikidata would need to adapt the concept of edit sessions [10].

Secondly, we deliberately restricted our analysis to a number of namespaces. We are, for example, aware of the importance of the Wikidata community namespace in defining properties. In this namespace, every contributor is able to propose new properties that are then discussed and finally approved by the community[15]. The content analysis of items required in this namespace goes beyond the scope of this research, but will be addressed in future work. In this context, it might be interesting to compare editorial processes in Wikidata with other projects, such as Semantic MediaWiki communities (cp. [12]).

Thirdly, we did not take into account the semantic context of an edit. The best way to understand the semantic contribution of edits is to match them with a RDF representation from Wikidata (as created by Erxleben *et al.* [8]). Strohmaier *et al.* [27] show that existing semantic relations in a structured knowledge base influence the way in which this knowledge base is edited. Context information such as this can also be used to reconstruct contributor sessions in order to determine semantic editing paths. This might also be interesting from a user interface design perspective, since Wikidata aims to allow everybody to participate, regardless of whether or not they are familiar with semantic technologies.

8. CONCLUDING REMARKS

Peer-production communities addressing the development of structured data have not as yet attracted much attention from the research community. Scholars focus primarily on communities in the area of open source software development and Wikipedia. Research on collaborative ontology engineering projects exists, but, so far, is discussed predominantly on the Semantic Web or in the Knowledge Representation research community. With the emergence of a new project in the Wikimedia ecosystem – Wikidata – the shadow existence of the collaborative construction of structured datasets may come to an end. Pre-existing software systems for collaborative ontology engineering require proficient contributor understanding of semantic web technologies and ontology languages. Wikidata provides the prototype of a system that allows even non-technical experts to create and manage semantic data. Wikidata could be the nucleus for a completely new type of system.

We might augment existing tool developments in the Semantic Web area as well as in Social Computing by bringing research on peer-production communities and collaborative ontology engineering projects more closely together. We have taken a first step in this direction with our research.

[15] http://www.wikidata.org/wiki/Wikidata:Property_proposal

By understanding the innermost functions, i.e. the participation infrastructure, of the Wikidata community, we can identify commonalities within peer-production communities and can transfer insights from one research area to the other. Nonetheless, our results are not conclusive and merely reflect a state in the development of Wikidata. As we have discussed in previous sections, the Wikidata community faces various challenges to come. It will be interesting to trace this development further, in order to verify to what extent our identified participation patterns remain constant or evolve over time.

9. ACKNOWLEDGMENTS

This work was partially supported by grant DFG MU 3146/1-1 from the German research foundation (DFG). The authors would like to thank the anonymous reviewers for their helpful and constructive comments, which greatly contributed to improving the final version of the paper. We would also like to thank Lukas Benedix for pre-processing the Wikidata data dump and Carola Zwick for her graphical design.

10. REFERENCES

[1] O. Arazy et al. Functional Roles and Career Paths in Wikipedia. In *Proc. CSCW 2015*.

[2] A. Ben-Hur et al. A Stability Based Method for Discovering Structure in Clustered Data. *Pacific Symposium on Biocomputing*, 7(6):6–17, 2002.

[3] K. Bollacker et al. Freebase: A Shared Database of Structured General Human Knowledge. In *Proc. AAAI 2007*.

[4] F. Bry et al. Semantic wikis: Approaches, applications, and perspectives. In T. Eiter and T. Krennwallner, editors, *Reasoning Web. Semantic Technologies for Advanced Query Answering*, pages 329–369. Springer, 2012.

[5] S. L. Bryant et al. Becoming Wikipedian: Transformation of Participation in a Collaborative Online Encyclopedia. In *Proc. GROUP 2005*.

[6] B. Butler et al. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proc. CHI 2008*.

[7] D. Cosley et al. How oversight improves member-maintained communities. In *Proc. CHI 2005*.

[8] F. Erxleben et al. Introducing Wikidata to the Linked Data Web. In *Proc. ISWC*, 2014.

[9] S. Falconer et al. An Analysis of Collaborative Patterns in Large-scale Ontology Development Projects. In *Proc. K-CAP 2011*.

[10] R. S. Geiger and A. Halfaker. Using Edit Session to Measure Participation in Wikipedia. In *Proc. CSCW 2013*.

[11] R. S. Geiger and A. Halfaker. When the Levee Breaks: Without Bots, What Happens to Wikipedia's Quality Control Processes? In *Proc. OpenSym 2013*.

[12] Y. Gil and V. Ratnakar. Knowledge Capture in the Wild: A Perspective from Semantic Wiki Communities. In *Proc. K-CAP 2013*.

[13] T. Gruber. Ontology. In L. Liu and M. T. Özsu, editors, *Encyclopedia of Database Systems*, pages 1963–1965. Springer US, 2009.

[14] A. Kittur et al. He says, she says: conflict and coordination in Wikipedia. In *Proc. CHI 2007*.

[15] T. Kriplean et al. Articulations of wikiwork: uncovering valued work in wikipedia through barnstars. In *Proc. CSCW 2008*.

[16] M. Krötzsch et al. Semantic Wikipedia. *Journal of Web Semantics*, 5(4):251–261, 2007.

[17] J. Lave and E. Wenger. *Situated Learning: Legitimate Peripheral Participation (Learning in Doing: Social, Cognitive & Computational Perspectives)*. Cambridge University Press, 1991.

[18] J. Liu and S. Ram. Who does what: Collaboration patterns in the wikipedia and their impact on article quality. *ACM Transactions on Management Information Systems (TMIS)*, 2(2):11, 2011.

[19] M. Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[20] A. Mockus et al. A case study of open source software development: the Apache server. In *Proc. ICSE 2000*.

[21] C. Müller-Birn et al. Work-to-rule: the emergence of algorithmic governance in Wikipedia. In *Proc. C&T 2013*.

[22] K. Nakakoji et al. Evolution patterns of open-source software systems and communities. In *Proc. IWPSE 2002*.

[23] F. Ortega and J. M. G. Barahona. Quantitative analysis of the wikipedia community of users. In *Proc. WikiSym 2007*.

[24] J. Preece and B. Shneiderman. The Reader-to-Leader Framework: Motivating Technology-Mediated Social Participation. *AIS Transactions on Human-Computer Interaction*, 1(1):13–32, 2009.

[25] G. Robles, J. M. Gonzalez-Barahona, and I. Herraiz. Evolution of the core team of developers in libre software projects. Technical report, Universidad Rey Juan Carlos (Spain), 2005.

[26] E. Simperl and M. Luczak-Rösch. Collaborative ontology engineering: a survey. *The Knowledge Engineering Review*, 29(01):101–131, 2013.

[27] M. Strohmaier et al. How Ontologies Are Made: Studying the Hidden Social Dynamics Behind Collaborative Ontology Engineering Projects. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2013.

[28] T. Tudorache et al. Supporting collaborative ontology development in protégé. In *Proc. ISWC 2008*.

[29] T. Tudorache et al. Using Semantic Web in ICD-11: Three Years Down the Road. In *Proc. ISWC 2013*.

[30] D. Vrandečić and M. Krötzsch. Wikidata: A Free Collaborative Knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[31] H. T. Welser et al. Finding social roles in Wikipedia. In *Proc. iConference 2011*.

[32] E. Wenger. Communities of Practice: Learning as a social system. *System Thinker*, 9(5):2–3, 1998.

[33] Y. Ye and K. Kishida. Toward an understanding of the motivation Open Source Software developers. In *Proc. ICSE 2003*.