Wikidatians Are Born: Paths to Full Participation in a Collaborative Structured Knowledge Base

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

Abstract

We investigated how participation evolves in Wikidata as its editors become established members of the community. Originally conceived to support Wikipedia, Wikidata is a collaborative structured knowledge base, created and maintained by a large number of volunteers, whose data can be freely reused in other contexts. Just like in any other online social environment, understanding its contributors' pathways to full participation helps Wikidata improve user experience and retention.
We analysed how participation changes in time under the frameworks of legitimate peripheral participation and activity theory. We found out that as they engage more with the project, "Wikidatians" acquire a higher sense of responsibility for their work, interact more with the community, take on more advanced tasks, and use a wider range of tools. Previous activity in Wikipedia has varied effects. As Wikidata is a young community, future work should focus on volunteers with little or no experience in similar projects and specify means to improve critical aspects such as engagement and data quality.

1. Introduction

The rise of Web 2.0 has facilitated the proliferation of open production communities (OPCs) [2]. Technology-mediated social participation [26] has been used extensively to enable the most diverse forms of collaborative creation process [37]. One of the newest and most interesting examples of OPCs is Wikidata, launched by the Wikimedia foundation in October 2012. This initiative leverages the "wisdom of the crowd" to create a knowledge graph (KG), or a structured knowledge base— that is, a collection of terms describing entities, classes of entities, their characteristics, and the relationships that hold between them [30]. KGs are useful because they make it easy to manage, access, and aggregate information in great detail. For example, a query against Wikipedia for all cities in France which hosted two *EURO*2016 football games or more will only return the expected results if this list has already been compiled (manually) from different articles (such as the article of *EURO*2016 or the ones for French cities). By comparison, a KG can answer this, and any other query, on the fly, as the information is available in structured, predictable form, which can be re-purposed and mixed without effort. A comparison with other KGs shows better the importance of Wikidata in this context. Wolfram Alpha is a question-answering system based on structured data curated by experts. It provides high-quality information, but cannot be used freely by other parties. By contrast, Wikidata's open licence allows anyone to share, reuse, and modify its data, without the need to specify any attribution [8]. This creates opportunities to develop a variety of data-centric applications on top of it, including question-answering systems similar to Wolfram Alpha. In addition, it is most likely less costly to maintain, as it relies on volunteer contributions. Another example is DBpedia, a central node of the Linked Open Data cloud [22]. Just like Wikidata, it does not restrict access and use, which makes it a popular source of data by many applications. However, DBpedia is "static", in the sense that its database is periodically automatically extracted from Wikipedia infoboxes using pre-defined mapping rules [22]; its users cannot directly modify it, which makes correcting errors a lengthy process. One of the advantages of Wikidata is its community of editors [32], who constantly update and fix inconsistencies, and are able to provide globally optimal solutions to problems, due to their diverse backgrounds [1].
Since Wikidata is completely community-driven, it can be successful only if it continues to grow and improve, both in terms of content and editorial practice. Among other things, Wikidata needs new editors to join its community and become experienced contributors. Interactions between novices and experts[1] have been shown to be important for the outcome of OPCs. However, the effects of adding new members of an established com-

[1]There is no established vocabulary to define users according their level of experience on OPCs. We primarily employed the terms novices and experienced contributors. Nevertheless, for stylistic reasons, we used sometimes synonyms, e.g. newcomers for the former and established or seasoned contributors for the latter.

munity can be mixed— while they help OPCs scale, improve their churn rates, and refresh their ideas, they can more easily deviate from accepted community norms, which may make existing members leave or engage less [29]. The role of novices changes as a whole across the life span of a community, as the importance of their contribution increases as the system matures [19]. How they evolve into seasoned contributors is another important aspect— OPCs' tend to have low entry and exit barriers [14], which means that their retention is critical to maintain a certain level of activity and keep the community alive. To offer the best community experience, both for its members and the platform as a whole, this evolution needs to be studied from a socio-technical point of view, including aspects such as community dynamics, and the interaction of individuals with the medium. Gaining insights over these aspects would also be beneficial to assess the crowd capital of Wikidata, i.e. the organisational-level resource generated by its community [27], but to the best of our knowledge they have not been investigated yet in this project. This paper seeks to fill this gap by analysing the pathways to full participation of Wikidata editors. Through interviews with "Wikidatians" we found out how their perception of their role, the community as a whole, and the tools they are using change over time. One element that distinguishes Wikidata from other KGs is its surrounding ecosystem, of which Wikipedia and other Wikiprojects are part. Our analysis revealed that people who only recently joined Wikidata via Wikipedia develop a sense of ownership towards the former even when they had previously considered themselves "Wikipedians". As they engage more with the project, their goals shift, they embrace more advanced tasks (such as adding large number of Items in batch, maintaining the knowledge base structure, and mentoring). Prior experience in Wikipedia helps newcomers form their expectations about the new environment and shape their motivations to join, but does not seem to affect their initial relationship with the community, or the way they use technology.

In the following, we first introduce Wikidata and its community. Subsequently, we present the methodology adopted and discuss the results, before wrapping up with a conclusion and ideas for future work.

2. About Wikidata

2.1 Wikidata as a KG

Wikidata uses a *property-value*-based model to organise and encode knowledge. It consists of *Items*, described by means of *Properties* that connect them to other Items or to values. Items correspond to entities, both concrete ones, such as *Mick Jagger*, and abstract ones, such as the class of all musicians. Items, Properties, and values form *claims*, which can be optionally enriched with *references* and *qualifiers*. Together, they are referred to as *statements*, which are the main building blocks of the KG. References specify the source of a claim, in line with Wikidata's aim to be a collection of knowledge from primary sources [32]. Qualifiers add context, such as the temporal validity of a statement.

The same style of interface is used to edit Items and Properties. Each of them has a dedicated Web page. Revisions range from editing human-readable language labels— Wikidata uses unique alpha-numeric language-independent identifiers for its entities —to the modification of a claim. Different types of edits require different levels of knowledge and skill [4]. In some cases, e.g. changing the value of the population Property of a city Item, it is enough to grasp the basics of the interface and be familiar with the factual knowledge. In others, such as introducing a new Property, which may apply to many Items, it demands some command of knowledge engineering and an understanding of Wikidata's knowledge model. Besides Item and Property pages, Wikidata includes pages for editing policies and for discussion or talk, which serve as communication channels for the community. Editors may use their own talk pages to present themselves or exchange messages with their peers. Each Item and Property has a discussion page as well. Finally, Wikidata does not follow the same consensus-based model known from Wikipedia. Contrasting claims about the same Items co-exist, granted by qualifiers and references, and users of the KG may decide which claims they consider valid or trustworthy.

2.2 Wikidata as a community

The interest of Wikidata lies also in the peculiarity of its knowledge engineering processes. Users are responsible for both Wikidata's instance and conceptual knowledge, i.e. the single entities it describes and the the underlying knowledge representation. This distinguishes Wikidata from previous examples of peer-production systems, e.g. Wikipedia, and of collective ontology engineering projects, such as the ones mentioned in [12], and situates it at the intersection of the two typologies [24]. Previous analogue examples, e.g. "semantic wikis" [15], which combined axioms to unstructured text, or Freebase [13] never attained a breadth and user base comparable to Wikidata.

Anyone can contribute to Wikidata. Its KG is created by machines (bots) and humans alike, who can either register or contribute anonymously. There are currently more than 100 thousand registered editors. However, the level of activity varies greatly: according to our observations, about 2% of the community is responsible for

almost 95% of all manual edits. As such, the community has a core-periphery structure common to several other OPCs, e.g. Free Open Source Software projects [9], or Wikipedia [7]. A similar distribution can be found also in the revision scope. Only $< 2\%$ has ever edited any Properties and Items functioning as classes [24], i.e. the elements constituting the structure of Wikidata's knowledge, coherently with what [15] observed in other semantic wikis. Finally, many of Wikidata's top members have been involved in Wikipedia before joining, often in administrative roles. A manual check of the activity history of the 50 most active Wikidata contributors showed that all but five of them have registered and have been highly active in Wikipedia long before starting in Wikidata. This strong link can be explained by the original role of Wikidata as a support project for Wikipedia.

3. Methodological approach

This study explores how early forms of participation in Wikidata change across editors' activity lifespan. Especially, it examines how the shift from more marginal to fuller forms of participation transforms volunteers' motivations and self-perception of their role, how socio-technical structures mediate social activity in Wikidata, and to what extent these are perceived as barriers for participation. This section describes the approach followed to investigate these aspects, in terms of choice of theoretical frameworks, data collection and analysis.

3.1 Theoretical framework

In order to understand the evolution of users within Wikidata, we framed our analysis into existing theories regarding how community members interact and how their roles change over the course of their contribution.
The *Usage Lifecycle* approach categorises users of a social Web platform according to their usage patterns, with four distinct types— interested, first-time, regular, and passionate users [25]. A small portion of community members moves from one stage to another as the level of engagement increases, in a process plagued by various hurdles and challenges, which may discourage them to continue. The *reader-to-leader* framework [26] studies IT-mediated social participation. It identifies four types of community activity: readers, contributors, collaborators, and leaders. Each of these is influenced by usability and sociability factors, and characterised by several dimensions: motivation, relationship with others, awareness of community norms, and use of the interface. Changes between groups may go back and forth non-linearly, and users might not reach the status of leaders.
These two frameworks were not suitable for our study. The User Lifecycle does not take into consideration the social context in which novices operate and it is intended to measure the effectiveness of Web sites in a practical, design-centric context. Rather than examine in detail the interface design and its impact on user engagement, we wanted to look at how editors use specific tools as their role and motivation change. The reader-to-leader framework is a somewhat better fit, as it describes non-linear participation pathways from a socio-technical point of view. However, we relied on a simpler way to distinguish different levels of participation (novices and experts), which we believe is more appropriate for a rather young and emerging community such as Wikidata.
Legitimate Peripheral Participation (LPP) aims to understand how learning takes place within communities of practice [34]. LPP focuses on environments where the learning process happens as part of a social construction, therefore its application goes beyond situations that have learning as their primary focus [5]. In LPP each activity is situated in a social context— the learning process is about becoming an "insider" of a community of practice, eventually fully able to behave and act as a community member [5], rather than about acquiring abstract knowledge or gaining skills in some practical area. As such, learning happens by taking on increasingly difficult tasks and following experienced practitioners in their daily activities, similar to an apprenticeship [34]. Through the relationship to a master and peers, the apprentice acquires the necessary abilities to be part of the community. This is explained by the concepts of *legitimacy* and *peripheral participation*. Legitimacy is crucial for learning and makes newcomers become recognised as aspirant members of the community of practice, with the corresponding rights and duties. Peripheral participation must be intended as one of multiple forms of participation with diverse levels of engagement within a social context and fundamental for the learning to occur [34], rather than as secondary with respect to a centre.
Another way to study online communities is *Activity Theory* (AT) [21]. Originating from Human-Computer Interaction research, AT sees activities as the socio-technical and normative context in which human actions take place. Individually, human actions can only be understood within the context of an activity, which is the minimal unit to bear intelligible meaning. [21] describes how actions develop continuously, but unevenly, and illustrate the core concept by analysing a group of primitive hunters beating bushes to frighten a prey. Their behaviour can be understood only by considering their activity as a whole, i.e. hunting, in which they cooperate with another group that catches the frightened game. This example may be used further to explain the different elements constituting the basic structure of an activity. A reciprocal relationship between the hunter, the

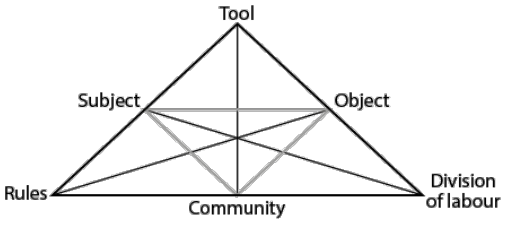

Figure 1. Activity Theory model [11] [21].

subject, and their prey, the *object*, exists, motivated by the hunter's will to obtain food and clothing material, the *outcome*, and mediated by the hunting *tools*. The transformation of an object into an outcome is the root of any activity [21]. To the triangle subject-tool-object traced so far, we add another vertex, the *community* to which a hunter belongs. As hunters take different roles, the *division of labour* mediates the relationship between the community and their prey. *Rules* and norms mediate the relationship of each hunter with the rest of the community. These relationships are shown in Figure 1.

Our study relies on a combination of LPP and AT as frameworks of analysis. Both have already been used to answer similar research questions. In a survey of readers of Wikipedia, [3] found a relationship between the levels of knowledge of the various editing features and the amount of time spent reading the site, with reading (or taking part without actively contributing or "lurking") found to be a key activity for newcomers to learn about the system. Additionally, as users spend more time reading discussions on the Wikipedia Talk pages, they transition to increasingly demanding tasks and gain knowledge about the operation of the site [3]. [31] studied users of a mobile tagging system and found that while experts participated at higher rates than novices, they expressed territoriality and were critical of the contributions made by others. It is thus important to ensure that novices are able to participate without a feeling of marginalisation. Using field experiments, [16] examines LPP in Wikipedia for increasing participation of readers and suggest that there are many potentially productive non-contributors that are unaware of their opportunities to contribute, which may be due to the fact that editing an encyclopedia is a demanding task. In another investigation of Wikipedia, [6] looked at the evolution of participation in the encyclopedia over time using LPP, and found how users move from a focus on individual articles to an overall concern for the site and community as a whole. Participants adopt new roles and goals, and begin to use a wider variety of tools available to them [6]. AT was also used in combination with LPP to help describe the evolution of these user roles in the community, providing a way to examine the different dimensions of human activity [6]. The progressive integration of users into the Wikipedia community was explained as a movement from peripheral to richer forms of participation [6]. LPP and AT were well suited to address our RQs. First, Wikidata's articulation of work entails diversified informal and formal user roles, which involve tasks requiring different levels of skills and responsibility. This matches well the LPP model of learning by doing and interacting with other members of a community. Second, AT encompasses the broader scope of the interactions between members within their community and directed to an object, contextually taking into consideration how these interactions are mediated by technical artifacts, community norms, and policies [37]. The combination of LPP and AT was thus suitable to address the research questions posed, as AT was used to code different aspects of user interaction within the socio-technical system of Wikidata, while LPP enabled to explain their evolution.

3.2 Methods and description of participants

We followed a qualitative approach to obtain rich insights into the editors' modes of participation in the Wikidata community of practice. The choice of semi-structured interviews ensured to cover the relevant aspects of the subject relationship with tools and community, while giving participants the freedom to provide new and unexpected information. We sought committed, experienced Wikidata editors. All respondents authored several thousands of revisions and had been active in Wikidata for several months. Rather than also contacting new users, we asked established users to look back at their first experiences and perceptions of their contribution to Wikidata. This provided a more coherent account of the transformation of users' perceptions from lateral to full participation. We recruited users by posting a message on the Wikidata Project Chat page, which is used to discuss general issues by contributors, and by direct messages on their personal pages. All participants opted to complete the interview via email.
The interview had three parts, each one containing questions that addressed aspects related to how the subject perceived itself, the community, and its overall work on Wikidata. The first questions asked them to reflect about their time as novices, whereas the second part focused on their present experience so that we could draw comparisons and identify changes in their behaviour. Finally, the third part enquired about the influence of their previous experiences on Wikipedia on their current activities. Of a total of 20 Wikidata users invited, eight eventually took part in the study, of which only seven submitted their responses on time to be included in this research. The interviewees, seven males and one female, were from six different countries. All were involved in other Wikimedia projects, mainly Wikipedia,

before starting to contribute to Wikidata, and six were highly active Wikipedians. All participants answered all the questions (19), where applicable, and generally provided several details about their experiences and impressions about their activity on Wikidata. The answers were coded by two researchers in two successive cycles. The number of participants was comparable to previous studies, e.g. [6], which had nine respondents. With regard to the conceptual saturation of the sample, the interview that has not been included in this study added little or no further insights to the ones already analysed. Preliminary discussions with the Wikidata team at Wikimedia Germany confirmed the impressions obtained from the interviews. This led us to conclude that the gathered responses were rich enough to draw sound conclusions.

4. Findings

We present our findings according to the elements of the AT model which they refer to. First, we describe the transformations undergone by the subject, in their perception of self, motivations, and goals. Second, we examine how the use of the interface and other tools evolves as novices gain experience within Wikidata. Finally, we cover the evolution of user perception of community, norms, and division of labour.

4.1 Subjects: how identity and goals change

Similar to the previous study using AT on Wikipedia [6], the data from our interviews suggests that subjects do not change the object they aim to transform, i.e. Wikidata's knowledge graph. What does change is the subjects' perception of their own identity as contributors to Wikidata, their motivations, and goals. These transformations are described in the ensuing sections.

4.1.1. Novices All respondents already had previous experience with Wikipedia when they started contributing to Wikidata. This affected their approach to the new project. The previously established Wikipedia editors continued to primarily identify as such and emphasised their being Wikipedians by pointing out the length of their contribution to the encyclopedia. Some of them underlined their growing detachment from Wikipedia.

> (P1) "I've been involved with Wikipedia and less with other WMF projects since December 2004. I've been actively involved with Wikimedia UK since 2013."
> (P3) "I've been a Wikimedian since 2008, the idea of Wikidata was around before that, so I knew about the project years before it was actually launched. I was mostly a Wikipedian at the time."
> (P5) "I started in 2001 on the English Wikipedia, in 2003 I was elected administrator and for a few years I was very active there. As time passed, school and work kept leaving me less and less time for Wikipedia."

Some respondents were Wikidata newcomers who have collaborated to Wikipedia, without being heavy contributors. They tended to define themselves more in relation to the new platform, e.g. as "simple participants" (P3). Interviewees were attracted to Wikidata for multiple reasons, somewhat intertwined with their Wikipedian identity. Whereas some contribute to Wikidata to "solve purely Wikipedia problems" (P3), the majority are attracted by the possibilities given by Wikidata's inherent features, which enable them to look beyond a merely "Wikipedia perspective" and think about it as a new project in its own right. Structured data is an important motivation, as it is easy to reuse and connect to other data sources. As well, there is an appreciation for the open nature of Wikidata (P5) and a willingness to identify with and contribute to its community goals (P1, P5).

> (P1) "Most rewarding is the sense of contributing structured data that can be reused in many ways."
> (P3) "I loved organising the human knowledge, structuring data, and make it usable and shared. I love that we can now write incredibly detailed queries and have decent answers because Wikidatians added this data."
> (P5) "I loved that it was built to be a proper semantic knowledge base and not just another database. At the beginning many people saw Wikidata as an extension to Wikipedia, but I wouldn't have bothered to contribute in that case. For me it has always been about aggregating all the data in the world in a semantic knowledge base that is easy to edit and easy to query. I loved that I could query it and retrieve anything that had been inserted and immediately reuse it. I loved that Wikidata is collecting data and identifiers from hundreds of databases and making everything connected."
> (P7) "I knew that my contributions to Wikidata would not only help users directly read and search the database, but they'd also improve computational models for anyone who wishes to build them."

As newcomers, interviewees were also attracted by Wikidata's simplicity and support for different languages, often in opposition to Wikipedia. Multilinguality allows people from all around the world to contribute to a common project (P5), while making one's contributions available to more people (P7). Some respondents were "fascinated by the simplicity of Wikidata in comparison to Wikipedia" (P4) or found contributing to the KG "a stress release to the disputes at Wikipedia" (P1), up to the point that technical discussions and intricacies connected to structuring data can be seen as a threat to this simplicity (P4). Wikidata's alleged simplicity may stem from is its liberal attitude concerning contrasting

statements and Property constraints. Wikidata allows contrasting statements to co-exist and to express different points of view and opinions. Combined with the absence of enforced Property constraints, this enables users to simply add information, either as statements or human-readable labels and descriptions, without dealing with complex policies or community norms. Returning to the distinction made in Section 2 between edits requiring different skill levels, [17] categorises activities in OPCs as *lightweight* and *heavyweight*. Lightweight contributions are loosely coordinated and require less effort, thus setting lower entry barriers. Heavyweight ones necessitate higher levels of coordination and are more subject to agreed policies and group consensus. Wikidata supports both lightweight and heavyweight edits, helping to motivate novices and influencing their goals. We have already mentioned that reading articles, or "lurking", has been considered a form of peripheral participation in Wikipedia [3]. None of the interviewees had a passive approach to Wikidata, stating instead to have started carrying out revisions straight away. This may offer some support to the findings of [10] for Wikipedia, where the biggest contributors began immediately with large edits. Interestingly, a participant remarked that "there was never anything refraining [him] from editing" (P5). The above mentioned relaxed approach of Wikidata to contributions may thus prove effective in driving engagement and retaining newcomers.

The interviewees reported to have had different behaviours as novices. Several participants started by editing Items related to a topic of interest. This theme can be either a broader, e.g. fictional works (P3), or a more circumscribed one, e.g. a country (P4) or its local administration (P2). This approach can coexist with a focus on particular scopes of revision, e.g. editing labels or descriptions (P1) or statements using a particular Property (P3). Finally, in line with their identity as former Wikipedians, some participants focused on edits with a direct impact on Wikipedia (P3, P7), considering Wikidata mainly as a support for the online encyclopedia.

4.1.2. Experts Participants began to identify themselves with the project after a certain number of edits. People who used to consider themselves Wikipedians built an identity around the Wikidata community. Wikidata has become their main focus now— they feel they are already "established contributors" (P1), some of them define themselves as "Wikidatians". This change in perception of self entails a different, wider, and higher-level range of responsibilities, which may translate into formal administrative roles.

> (P3) "I'm a Wikidatian, I add information, references, labels, etc. I add and clean it. I'm also a sysop [*i.e. a user with administrative privilege*], and as such I do technical edits: deletions when needed, blocking vandals, cleaning up vandalism. I learned about semantic web, data, ontologies, for Wikidata, I didn't know anything about that before. I had no idea what a database could be used for and that I could find it interesting!"
> (P5) "As an administrator, I feel responsible for the protection and advancement of the project. It definitely is [*my main project*] and I love it."
> (P7) "Wikidata is my main project. After a year editing both it and Wikipedia, I was about half of my time on each. In the second year, I became mainly a Wikidata contributor."

The identity mutation does not appear to bring radical changes to editors' motivations. Some clearly stressed that their motivation has not undergone any transformation since they began contributing (P2, P3, P5). For the majority, contributing to a KG that can be easily shared and reused by many people is a primary reason also as established contributors. One respondent "loves to see [her] work used" (P3), another one likes "to think about what will people do with the data [he] inserts, and it makes [him] feel good" (P5). Still for another one learning and acquiring knowledge are the motivation behind his contribution to Wikidata (P2). The motivations hitherto noted may be explained by recurring to the categorisation of member motivations within virtual communities in [23]. Altruism and collaboration can enclose the attention given by both newcomers and experienced participants to reuse and share data; the stress on learning and gaining new knowledge can instead be related to wisdom and knowledge. Altruism, collaboration, and knowledge are among the motivations for taking part in both wikis and knowledge bases, whereas wisdom, i.e. obtaining new information and expertise, is associated only to the latter [23]. Ease of use becomes secondary for experienced users. First, novices consider simplicity as in contrast to their present identity as Wikipedians. Once they become more acquainted with the new system and start regarding themselves as fully integrated participants, they do not see anymore Wikidata's features in opposition to the previous system. Second, simplicity increases user satisfaction, but is not a motivation in itself [20]. As such, the possibility to perform low-effort contributions may again strengthen novices' motivation, thus being more likely to retain them afterwards.

Users' goals change, in spite of the small shift in motivation. Those who initially focused on adding links between different Wikipedia language versions started editing Items' statements and labels (P1, P3). Those who had been editing statements since their first approach generally moved to higher-responsibility tasks. These include maintenance-related tasks, such as identi-

fying and fixing Property constraint violations (P2), and addressing quality problems to help Wikidata grow and serve as a base for the development of other projects (P7). Moreover, many Wikidatians work on revisions related to Wikidata's knowledge structure, proposing new Properties and Property constraints, and designing the structure of the Wikidata ontology. This is not formally designed as such, but results from the relationships between Properties and Items. This appears to reflect the findings from Wikipedia in [6] which showed how users move from focusing on editing individual articles to taking on greater responsibility for the system as a whole.

4.2 Tools: evolution in the use of the interface

Together with a change in identity and goals, the evolution to established Wikidatians involves the use of a more diverse set of features of the Wikidata interface. Following, we describe this transformation.

4.2.1. Novices We have noted that interviewees performed edits from the start of their collaboration, although possibly with different goals. Items and Properties can be easily modified through a simple interface. We refer to the ensemble of functionalities on a page that are used to edit its related Item or Property elements, e.g. labels and statements, as *basic interface*, leaving out tools used to manage edits or automated editing tools (Figure 2). This is the most direct way to perform edits on Wikidata and was used by several interviewees as novices (P4, P5, P6, P7), who noted that previous experience with Wikipedia was not helpful for this type of edit. Familiarity with the MediaWiki software— employed in Wikipedia and in many other wikis, including Wikidata —helped with other parts of the interface.

> (P5) "Wikidata has a different interface from other wikis (including Wikipedia), but it is very simple so I didn't need to know much. My experience with Wikipedia was useful mostly for discussions (i.e. wikitext syntax) and for the general interface of the MediaWiki software, e.g. where to look for page history and user contributions, what to find in the preferences."

The low entry barrier to Wikidata enables peripheral participation of newcomers, similarly to what happened in the case studies in [34]. No formal barriers restrain new Wikidatians from trying out other interface features. Our interviewees' responses suggest that previous experience within Wikimedia gave them the necessary confidence to use other tools once they began editing. One participant experimented with different tools as a novice (P2), while others used a Watchlist (P1, P5)— a tool that "bookmarks" pages of interest and sends notifications when they are modified —or communication channels, e.g. talk pages (P3). Still, they mostly used the basic

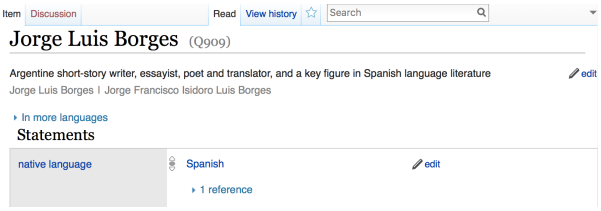

Figure 2. A Wikidata Item: by clicking on the *edit* link, users can modify the related element. On the top, the links to the Discussion page, its Revision history, and the search box. The star allows to add the Item to one's Watchlist. The same layout is used for Properties.

interface in their first encounters with Wikidata.

4.2.2. Experts Experienced users appear to continue to carry out revisions through the basic interface, which enables them to modify single Items that they are interested about. In addition, they more often integrate this with other ways to interact with the system, such as automatic editing tools (AETs). [6] rely on the concept of *zone of proximal development* formulated by Vygotsky [33] to explain how users' awareness of Wikipedia interface increases as they move towards full participation. The ZPD encloses the set of tasks that a person is unable to do without help and guidance from someone more experienced and knowledgeable. No such role of more established users was mentioned by the interviewees. Their awareness of tools other than basic editing interface, talk pages, and Watchlist varies. While one respondent stated "as I started, I knew nothing about [*automatic editing*] tools, so I did not use them" (P4), the others were aware of their existence, but either felt they did not need AETs for their purposes or became more acquainted with them by means of trial and error. When they were novices they appeared to not rely on other users to gain confidence in the use of the new platform. Instead, the adoption of new tools is interconnected with the evolution of goals. Established Wikidatians seem to find it more convenient "to use scripts and tools to facilitate [*their*] work" (P3) to be able to quickly carry out more edits, patrol against vandalism, and perform maintenance work. As experienced users, the interviewees described how they use more regularly the Watchlist, which allows them to keep track of the pages they are interested in. Their feeling of accrued responsibility towards Wikidata as a whole leads them to be willing to control a larger number of pages. To be aware of quality issues, they also use database reports. These are generated by programs that find, among other things, missing labels or statements and Property constraints violation.
A large number of AETs is available for Wikidata [35]. These tools allow editors to submit edits in batch, e.g.

to add several statements at the same time. They are employed by the interviewees to carry out a much bigger volume of edits, e.g. Quick Statements [35] allows to add several *subject-property-value* triples simultaneously, using a csv file. AETs also select Items using a particular Property or in a determined relation with other Items, e.g. all Items instances of *writer*, thus allowing one to find entities which meet their interest. This has been reported as an issue by several respondents, e.g. for P4 an AET, the Terminator, "solves one important problem: how to start, how to find Items?". AETs may also enhance user motivations: one editor (P3) was motivated by the gamification approach of the Wikidata game [35]. Respondents were more aware of the opportunities for engagement on talk pages as their experience increased. Community talk pages, e.g. Project Chat or Property proposal pages, become more of a space to keep oneself informed about and to discuss new developments and issues of Wikidata. This suggests that as with [3]'s study of Wikipedia, increased time spent on these will indicate a transition to more demanding tasks across the site. Frequenting community talk pages does not correspond to an increased use of User talk pages, though. Almost all interviewees host some content on their User talk page, but they provide minimal personal information.

4.3 Rules, community, and division of labour: the perception changes

In AT, rules and norms mediate subjects' interaction with the community. In turn, community action towards the object takes place through the assignment of different roles to its members. Due to this strict interconnection between rules, community, and division of labour, we discuss how they are transformed in the same section, adopting the approach followed by [6].

4.3.1. Novices Previous experience on Wikipedia made users aware of Wikidata's collaborative nature, with a community and its own policy, before they joined. This was judged to have a positive effect, as interviewees "knew what to expect from the community and the project's policies" (P5). For one, it was a hindrance though, given the differences between the two projects.
Wikidata's liberal approach with regard to contrasting statements and Property constraints is part of a more general effort to keep the number of policies on Wikidata low. Norms are in place to specify what the different elements constituting Items and Properties are and to regulate them. Knowledge Engineering principles are followed to structure Wikidata's knowledge. It is possible that to not "scare newcomers away", none of these rules are presented on the Item pages, and the help page that introduces users to editing in Wikidata [36] emphasises only that Items must be *Notable*, *Unique*, and *Linked*. Recalling the importance given to simplicity by newcomers, the interviewees seemed to know at least some of the principal Wikidata policies and valued that these are clear and not overnumbered, although one participant was "not always clear about Wikidata notability vis-a-vis Wikipedia" (P7). At the same time, interviewees do not mention knowledge engineering principles to be important during their first times on Wikidata.
As novices, respondents tended to have scarce relationships with the community, with the exception of one user involved in local promotion activities. Contacts with other users were minimal and mostly related to comments— which are appreciated —about the revisions made. This seems to contradict previous observations about open source software ecosystems. According to [18], users with experience and skills matured within a software ecosystem appear to focus more on managing and mentoring other members of the community when migrating to another project, rather than providing central code contributions. The differences between Wikipedia and Wikidata are more substantial than those between the projects in [18]. Despite belonging both to the Wikimedia ecosystem, they use different content models, have different interfaces, and different policies. Coherently with their sporadic relationships with the community, novices do not take formal roles, e.g. as administrators. Their focus is still the production of knowledge, by editing Items of their interest, and the exploration of the possibilities offered by the system.

4.3.2. Experts Once established users, the interviewees had more in depth knowledge of policies, as they now express judgements and ideas to change them. A step forward with respect to when they were newcomers is also the attention given to the knowledge engineering concepts necessary to build the structure of Wikidata's KG. Awareness of these concepts grew enough to enable them to propose new properties and create ontologies. Some concepts return several times in the interviewees' words, such as the difference between the Properties *instance of* and *subclass*, which appear to be commonly mistaken [4] and causes one respondent to complain about novices: "mostly what annoys me is people who don't understand how ontologies work, who confuse instance and subclass" (P3). As discussed earlier, in other systems there have been cases of experienced users similarly expressing territoriality, potentially intimidating newcomers, and this should be avoided [31].
Interactions with the community become more frequent for the users in our study. Now that they are experienced users on the site they feel "part of something big" (P3), have more familiarity with other users, and take part in discussions about diverse topics related to

the maintenance and expansion of the knowledge base. There are differences, however. One user is highly active, both on- and offline, as she organises events presenting Wikidata to the public, while another seldom contacts with other users. Yet another enjoys Wikidata because relationships with the community are not necessary. New ways to interact with other members reflects also an increased feeling of responsibility towards the whole project and the new role, formal or informal, taken by fully integrated users. As they became more experienced, many interviewees began mentoring other new members. They would welcome them and explain Semantic Web concepts, suggest changes, and discuss their decisions. One user (P4) stated: "I'm also a veteran Wikidatian now, so I welcome new users, I explain rules and the basis of the Semantic Web". Mentoring novices is not the only role assumed by seasoned users. Whether they fulfil a formal role in Wikidata hierarchy, e.g. administrators, or not, they generally undertake similar tasks. They oversee knowledge base quality, by fighting vandalism, fixing errors, and integrating information. They also undertake higher-level tasks, curating knowledge structure and setting up new Wikiprojects.

5. Limitations and future work

This study provides insights into how forms of user participation change over time within Wikidata and explores how this is influenced by previous activity in a "sister" project. However, we must point out some limitations with our approach. We selected a number of already established contributors to Wikidata and interviewed them to understand how their experiences had changed, rather than sourcing the views of complete newcomers. Participants were all previously involved in Wikipedia to some degree, which we have seen to be extremely common among committed Wikidata users with large number of revisions. This might signify that other users chosen according to the same criteria would have generated a similar story, but this could change in the future. In particular, it is likely that as Wikidata grows, so too will the number of editors coming to it without past experience even on Wikipedia. Their experiences may thus diverge from those highlighted in this paper. Future research could examine how completely new editors' behaviour changes as they become integrated into the community, to understand the process of novices establishing themselves as valuable contributors. Besides, the strong connection with the existing Wikipedia community had likely facilitated the construction of a crowd for Wikidata. This is crucial to generate crowd capital [28] and deserves to be addressed in further studies.

6. Conclusions

We investigated how newcomers become established members of the Wikidata community, carrying out our analysis within the frameworks of Legitimate Peripheral Participation and Activity Theory. We performed semi-structured interviews with committed members of Wikidata to understand how their identity and motivations, use of the interface, and perception of the community changed along with their activity on the platform. All the interviewees had previous experiences on Wikipedia, which allowed to gain insights also on transformation of user activity and community roles across platforms within the same online ecosystem. Along with their activities on Wikidata, novices acquired a feeling of identity with the project, even when they are former Wikipedians. For many, the initial motivation to join stemmed from the potentialities of structured data, combined with the alleged simplicity of the system, in contrast to Wikipedia. While motivation did not change over time, the respondents' goals and use of the interface did. Experienced users focus more on maintenance tasks, mass edits, and work on the structural elements of Wikidata's knowledge, using not only the basic editing interface, as they did as novices, but also AETs, to perform a larger number of revisions. As interviewees developed established relationships with the community, they used more often the communication channels, to offer advice and comments to the wider community. Table 1 presents the insights gained from this work.
Like colonisers of a new land, Wikidatians adapt their old habits and customs to a new environment. A generation of native Wikidatians may be already out there and will be the object of future research.

7. Acknowledgement

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).

8. References

[1] A. Afuah and C. L. Tucci, "Crowdsourcing as a solution to distant search," *Academy of Management Review*, vol. 37, no. 3, pp. 355–375, 2012.

[2] C. Aguiton and D. Cardon, "The strength of weak cooperation: An attempt to understand the meaning of web 2.0," *Available at SSRN 1009070*, 2007.

[3] J. Antin and C. Cheshire, "Readers are not free-riders: reading as a form of participation on Wikipedia," in *Proc. of the 2010 ACM conference on Computer supported cooperative work*. ACM, 2010, pp. 127–130.

[4] F. Brasileiro, J. P. A. Almeida, V. A. Carvalho, and G. Guizzardi, "Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata," in *Proc. of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 975–980.

Table 1. Summary of findings, categorised around the high-level category of results.

Finding Category	Finding
Attractions of Wikidata	– Contributing structured data, spreading and sharing knowledge. – Characteristics of Wikidata: simplicity; support for multiple languages; low entry barrier.
Motivation and perception of self	– Identity changes as level of responsibilities increases and admin roles taken on. – Motivation does not change inline with identity.
Roles	– General progression to higher responsibility tasks: creation of conceptual knowledge; quality control and maintenance. – Mentoring newcomers: introducing them to the community and to Semantic Web concepts.
Rules and norms	– Knowledge of Wikipedia increases awareness of norms. – Limited community relationships both as novices and, on a smaller scale, as experts.

[5] J. S. Brown and P. Duguid, "Organizational learning and communities-of-practice: Toward a unified view of working, learning, and innovation," *Organization science*, vol. 2, no. 1, pp. 40–57, 1991.

[6] S. Bryant, A. Forte, and A. Bruckman, "Becoming Wikipedian: transformation of participation in a collaborative online encyclopedia," *Proc. of the 2005 international ACM SIGGROUP conference on Supporting group work*, pp. 1–10, 2005.

[7] B. Collier and R. Kraut, "Leading the Collective: Social Capital and the Development of Leaders in Core-Periphery Organizations," *Collective Intelligence*, 2012.

[8] C. Commons. (2016, June) CC0 1.0 Universal (CC0 1.0). Public Domain Dedication. [Online]. Available: https://creativecommons.org/publicdomain/zero/1.0/

[9] K. Crowston, K. Wei, Q. Li, and J. Howison, "Core and periphery in free/libre and open source software team communications," in *Proc. of the 39th Annual Hawaii International Conference on System Sciences*, vol. 6. IEEE, 2006, pp. 118a–118a.

[10] S. Dejean and N. Jullien, "Big from the beginning: Assessing online contributors behavior by their first contribution," *Research Policy*, vol. 44, no. 6, pp. 1226–1239, 2015.

[11] Y. Engeström, R. Miettinen, and R.-L. Punamäki, *Perspectives on activity theory*. Cambridge University Press, 1999.

[12] S. Falconer, T. Tudorache, and N. F. Noy, "An analysis of collaborative patterns in large-scale ontology development projects," *Proc. of the sixth international conference on Knowledge capture - K-CAP '11*, p. 25, 2011.

[13] M. Färber, B. Ell, C. Menne, and A. Rettinger, "A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata and YAGO," *Semantic Web*, vol. 1, pp. 1–5, 2015.

[14] A. Forte and C. Lampe, "Defining, Understanding, and Supporting Open Collaboration: Lessons From the Literature," *American Behavioral Scientist*, vol. 57, pp. 535–547, 2013.

[15] Y. Gil and V. Ratnakar, "Knowledge capture in the wild: a perspective from semantic wiki communities," in *Proc. of the seventh international conference on Knowledge capture*. ACM, 2013, pp. 49–56.

[16] A. Halfaker, O. Keyes, and D. Taraborelli, "Making peripheral participation legitimate: reader engagement experiments in Wikipedia," in *Proc. of the 2013 conference on Computer Supported Cooperative Work*. ACM, 2013, pp. 849–860.

[17] C. Haythornthwaite, "Crowds and communities: Light and heavyweight models of peer production," in *Proc. of the 42th Annual Hawaii International Conference on System Sciences*. IEEE, 2009, pp. 1–10.

[18] C. Jergensen, A. Sarma, and P. Wagstrom, "The Onion Patch: Migration in Open Source Ecosystems," in *Proc. of the 19th ACM SIGSOFT Symposium and the 13th European Conference on Foundations of Software Engineering*, ser. ESEC/FSE '11. New York, NY, USA: ACM, 2011, pp. 70–80.

[19] A. Kittur, E. Chi, B. A. Pendleton, B. Suh, and T. Mytkowicz, "Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie," *World Wide Web*, vol. 1, no. 2, p. 19, 2007.

[20] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl, *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[21] K. Kuutti, "Activity theory as a potential framework for human-computer interaction research," in *Context and consciousness*. Massachusetts Institute of Technology, 1995, pp. 17–44.

[22] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, "DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[23] T. D. Moore and M. A. Serva, "Understanding member motivation for contributing to different types of virtual communities: a proposed framework," in *Proc. of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*. ACM, 2007, pp. 153–158.

[24] C. Müller-Birn, B. Karran, J. Lehmann, and M. Luczak-Rösch, "Peer-production system or collaborative ontology engineering effort: What is wikidata?" in *Proc. of the 11th International Symposium on Open Collaboration*. ACM, 2015, p. 20.

[25] J. Porter, *Designing for the Social Web*. Peachpit Press, 2010.

[26] J. Preece and B. Shneiderman, "The reader-to-leader framework: Motivating technology-mediated social participation," *AIS Transactions on Human-Computer Interaction*, vol. 1, no. 1, pp. 13–32, 2009.

[27] J. Prpic and P. Shukla, "Crowd science: Measurements, models, and methods," in *Proc. of the 49th Annual Hawaii International Conference on System Sciences*. IEEE, 2016, pp. 4365–4374.

[28] J. Prpić, P. P. Shukla, J. H. Kietzmann, and I. P. McCarthy, "How to work a crowd: Developing crowd capital through crowdsourcing," *Business Horizons*, vol. 58, no. 1, pp. 77–85, 2015.

[29] Y. Ren, R. Kraut, and S. Kiesler, "Applying common identity and bond theory to design of online communities," *Organization studies*, vol. 28, no. 3, pp. 377–408, 2007.

[30] S. Staab and R. Studer, *Handbook on ontologies*. Springer Science & Business Media, 2013.

[31] J. Thom-Santelli, D. Cosley, and G. Gay, "What do you know?: experts, novices and territoriality in collaborative systems," in *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1685–1694.

[32] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledge base," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[33] L. S. Vygotsky, *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

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

[35] Wikidata. (2016, May) External Tools. [Online]. Available: https://www.wikidata.org/wiki/Wikidata:Tools/External_tools

[36] ——. (2016, May) Help:items. [Online]. Available: https://www.wikidata.org/wiki/Help:Items

[37] P. Ziaie, "A Model for Context in the Design of Open Production Communities," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 29:1–29:29, Nov. 2014.