Proceedings of the 17th International Symposium on Open Collaboration

September 15–17, 2021
Online

General Chair	Gregorio Robles, Universidad Rey Juan Carlos, Spain
Program Chairs	Javier Arroyo, Universidad Complutense de Madrid, Spain Ann Barcomb, University of Calgary, Canada Kuljit Kaur Chahal, Guru Nanak Dev University, India Sulayman Sowe, Universität Bayreuth, Germany
Journal First Chair	Xiaofeng Wang, Free University of Bolzen-Bolzano, Italy
Sponsors	Universidad Rey Juan Carlos
In-cooperation	ACM SIGSOFT ACM SIGWEB

The Association for Computing Machinery
1601 Broadway, 10th Floor
New York, New York 10019, USA

ACM ISBN: 978-1-4503-8500-8

General Chair's Foreword

Welcome to OpenSym 2021, the 17th International Symposium on Open Collaboration which is organised 15th to 17th September. Given the still on-going global COVID-19 crisis, we are organizing this year's OpenSym again virtually and not in Madrid as it was originally planed.

Nonetheless, I am very pleased to host the premier conference on open collaboration research and practice and thereby promote stimulating discussions and dissemination of results in many areas of open collaboration, including open source, open data, open science, open education, wikis and related social media, Wikipedia, and IT-driven open innovation research.

> **Open collaboration** is egalitarian (everyone can join, no principled or artificial barriers to participation exist), meritocratic (decisions and status are merit-based rather than imposed) and self-organizing (processes adapt to people rather than people adapt to pre-defined processes)[1].

Many people have contributed greatly to make OpenSym 2021 happen and we owe them a great deal of thanks.

We are very grateful to all the researchers and practitioners that contributed to OpenSym 2021. The four program co-chairs, Ann Barcomb, Javier Arroyo, Kuljit Kaur Chahal, and Sulayman Sowe have made considerable contributions in time and effort for the conference program, especially in this situation, since we had to move a traditional conference to a virtual environment.

Sincere thanks go to Simon Butler for his considerable efforts as proceedings chair, to Antonio Balderas for having website, and social media always up-to-date, Amit Kumar Verma, Hernan Astudillo, Pablo Cruz Navea, Stefan Gruner, and S. R. S. Iyengar, our Regional Publicity co-chairs, and to Jesús M. González Barahona for his commitment as organisation chair. A special thanks goes to Dirk Riehle for his support and encouragement! We also want to thank all members of the Program Committee, the external reviewers, and all participants that contributed to OpenSym 2021.

Finally, we are also very grateful for the financial support the Universidad Rey Juan Carlos.

Without the aforementioned valuable contributions, OpenSym 2021 would never have happened!

Gregorio Robles
General Chair, OpenSym 2021

[1] https://opensym.org/about-us/definition/

Program Chairs' Foreword

We welcome you to the 17th International Symposium on Open Collaboration (OpenSym 2021), held online on 15–17 September 2021. As a premier conference in the field, OpenSym 2021 provides an excellent forum for reporting the latest developments on open collaboration research and practice, including open source, open data, open science, open education, wikis and related social media, Wikipedia, and IT-driven open innovation research.

In 2020, OpenSym, like many other conferences, was forced to make a sudden shift to a virtual event due to the COVID-19 pandemic. In 2021, we planned a hybrid conference, which would be held both online and in Madrid. However, the circumstances forced us again to celebrate it only online.

We have had our ideas about conferences challenged and reshaped in the last year and a half, and although there have been many difficulties, we have also had an opportunity to reflect on how events can be organized to address the differing needs of a diverse group of participants. We are inspired by the way geographically distributed collaboration teams work, and hope that the OpenSym conference can play a part in sharing knowledge of the many different ways communities can come together.

We are pleased to present the proceedings of the conference as its published record. We received 24 submissions and we selected 13 Full Research Papers, 1 Doctoral Consortium Paper and 3 Experience Reports. Research papers followed a double-blind peer review and are published as conference proceedings by the ACM, while the papers from the other tracks were single-blind peer-reviewed and are included in non-archival companion proceedings.

The conference program represents the efforts of many people. We want to express our gratitude to the members of the Program Committee and the external reviewers for their hard work in reviewing the submissions. The conference chair, Gregorio Robles also helped us in many ways, for which we are grateful. The paper submission and reviewing process was managed using the EasyChair system. We also acknowledge the fantastic work that Simon Butler, our Proceedings Chair, did in managing the conference proceedings. Finally, the conference would not be possible without the excellent papers contributed by authors. We thank all the authors for their contributions and their participation in OpenSym 2021! We feel honoured and privileged to serve as Program Chairs for the conference and hope that this program will further stimulate exciting research in all areas of open collaboration.

Javier Arroyo, Ann Barcomb, Kuljit Chahal, and Sulayman Sowe
Program Co-Chairs, OpenSym 2021

Organization

General Chair	Gregorio Robles, Universidad Rey Juan Carlos, Spain
Program Co-Chairs	Javier Arroyo, Universidad Complutense de Madrid, Spain Ann Barcomb, University of Calgary, Canada Kuljit Kaur Chahal, Guru Nanak Dev University, India Sulayman Sowe, Universität Bayreuth, Germany
Journal First Chair	Xiaofeng Wang, Free University of Bozen-Bolzano, Italy
Virtualization Chair	Jesús M. González Barahona, Universidad Rey Juan Carlos, Spain
Web and Social Media Chair	Antonio Balderas, Universidad de Cádiz, Spain
Steering Committee Chair	Dirk Riehle, Friedrich-Alexander University Erlangen-Nürnberg, Germany
Regional Publicity Chairs	Hernan Astudillo, Universidad Técnica Federico Santa María, Chile Stefan Gruner, University of Pretoria, South Africa S. R. S. Iyengar, Indian Institute of Technology Ropar, India Amit Kumar Verma, Indian Institute of Technology Ropar, India Pablo Cruz Navea, Universidad Técnica Federico Santa María, Chile
Proceedings Chair	Simon Butler, University of Skövde, Sweden

Program Committee

Oliver Baumann	University of Bayreuth, Germany
Hadj Bourdoucen	Sultan Qaboos University, Oman
Anamika Chhabra	Indian Institute of Technology Ropar, India
Kevin Crowston	Syracuse University, United States of America
Minh-Son Dao	National Institute of Information and Communications Technology, Japan
Henry Edison	The Mærsk McKinney Møller Institute, University of Southern Denmark, Denmark
Kristofer Erickson	University of Leeds, United Kingdom
Raula Gaikovina Kula	Nara Institute of Science and Technology, Japan

Matt Germonprez	University of Nebraska at Omaha, United States of America
Nikolay Harutyunyan	Friedrich-Alexander University Erlangen-Nürnberg, Germany
Benjamin Mako Hill	University of Washington, United States of America
Sudarshan Iyengar	Indian Institute of Technology Ropar, India
Nicolas Jullien	iSchool, Télécom Bretagne, LEGO-M@rsouin, France
Sherlock Licorish	University of Otago, New Zealand
Amanda Menking	University of Washington, United States of America
Reed Milewicz	Sandia National Laboratories, United States of America
Eunyoung Moon	Korea Advanced Institute of Science and Technology, South Korea
Claudia Müller-Birn	Freie Universität Berlin, Germany
Felipe Ortega	GSyC, ETSIT, Universidad Rey Juan Carlos, Spain
Lydia Pintscher	Wikimedia Deutschland, Germany
Mikko Rajanen	University of Oulu, Finland
David Rozas	Universidad Complutense de Madrid, Spain
Daniel Russo	Department of Computer Science, Aalborg University, Denmark
Giedre Sabaliauskaite	Coventry University, United Kingdom
Mirco Schönfeld	University of Bayreuth, Germany
Ioannis Stamelos	Aristotle University of Thessaloniki, Greece
Jose Teixeira	Åbo Akademi University, Finland
Xiaofeng Wang	Free University of Bozen-Bolzano, Italy
Astrid Wichmann	Ruhr Universität Bochum, Germany
Eva Zangerle	University of Innsbruck, Austria

Sponsors

Universidad
Rey Juan Carlos

In-cooperation

acm In-Cooperation

SIGSOFT
SPECIAL INTEREST GROUP ON SOFTWARE ENGINEERING

sigweb

Table of Contents

The platform belongs to those who work on it! Co-designing worker-centric task distribution models

David Rozas
drozas@ucm.es
Universidad Complutense de Madrid
Madrid, Spain

Jorge Saldivar
jasaldivar@ucm.es
Universidad Complutense de Madrid
Madrid, Spain

Eve Zelickson
eve@datasociety.net
Data & Society
New York City, United States of America

ABSTRACT

Today, digital platforms are increasingly mediating our day-to-day work and crowdsourced forms of labour are progressively gaining importance (e.g. Amazon Mechanical Turk, Universal Human Relevance System, TaskRabbit). In many popular cases of crowdsourcing, a volatile, diverse, and globally distributed crowd of workers compete among themselves to find their next paid task. The logic behind the allocation of these tasks typically operates on a "First-Come, First-Served" basis. This logic generates a competitive dynamic in which workers are constantly forced to check for new tasks.

This article draws on findings from ongoing collaborative research in which we co-design, with crowdsourcing workers, three alternative models of task allocation beyond "First-Come, First-Served", namely (1) round-robin, (2) reputation-based, and (3) content-based. We argue that these models could create fairer and more collaborative forms of crowd labour.

We draw on Amara On Demand, a remuneration-based crowdsourcing platform for video subtitling and translation, as the case study for this research. Using a multi-modal qualitative approach that combines data from 10 months of participant observation, 25 semi-structured interviews, two focus groups, and documentary analysis, we observed and co-designed alternative forms of task allocation in Amara on Demand. The identified models help envision alternatives towards more worker-centric crowdsourcing platforms, understanding that platforms depend on their workers, and thus ultimately they should hold power within them.

CCS CONCEPTS

• **Human-centered computing** → *Collaborative and social computing*; **Empirical studies in collaborative and social computing**;

KEYWORDS

crowdsourcing, digital labour, distribution of value, future of work, human computation, platform economy, task allocation, worker-centric platforms

ACM Reference Format:
David Rozas, Jorge Saldivar, and Eve Zelickson. 2021. The platform belongs to those who work on it! Co-designing worker-centric task distribution models. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3479986.3479987

1 INTRODUCTION

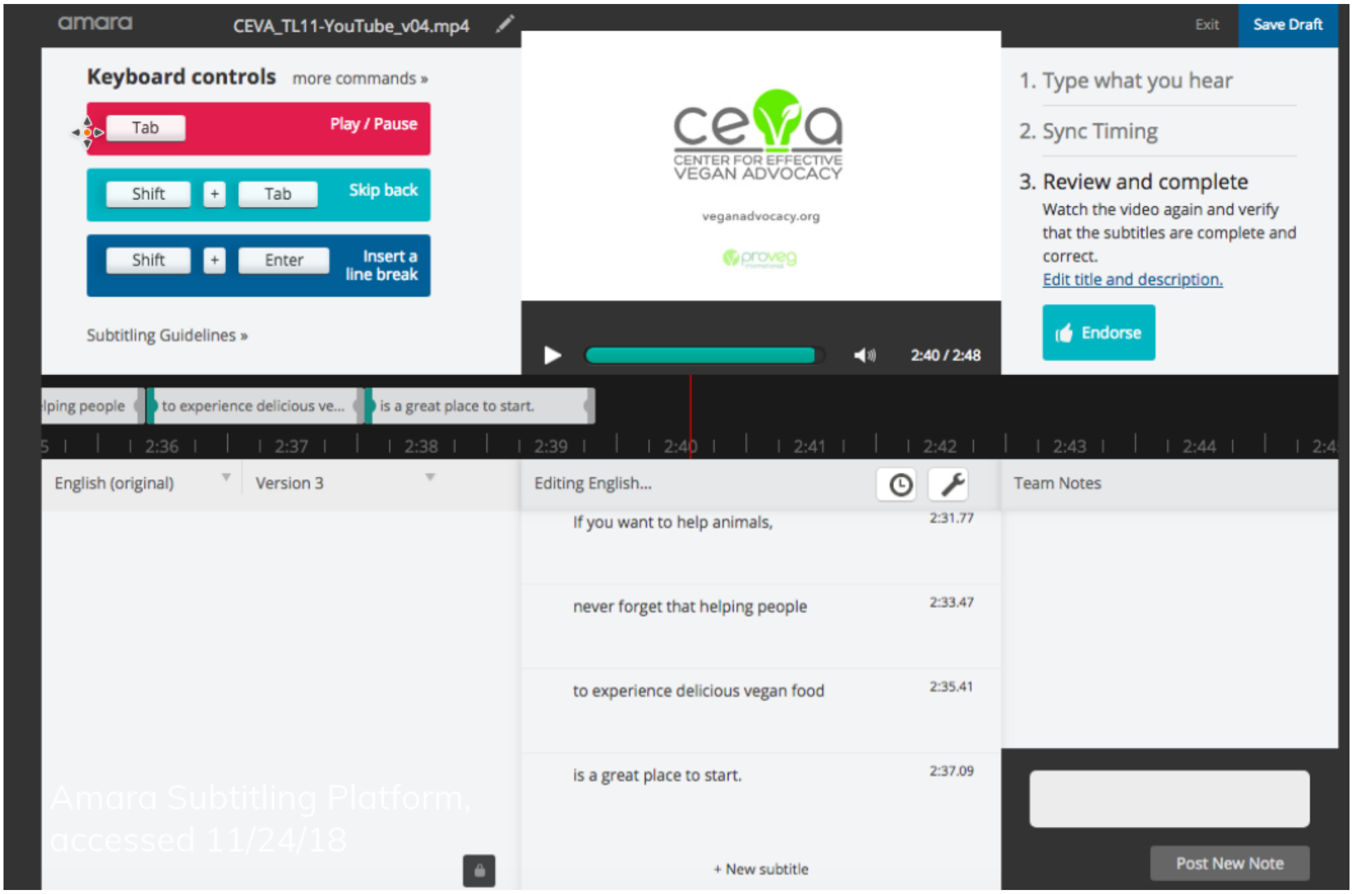
Contemporary working practices are changing and digital platforms are increasingly mediating our day-to-day work. In this scenario, crowdsourced forms of labour are progressively gaining importance, and large corporations such as Amazon and Microsoft are entering the field. Amazon Mechanical Turk (AMT), Universal Human Relevance System (UHRS), and TaskRabbit are all examples of market-driven crowdsourcing platforms. These platforms operate as labour marketplaces for businesses to outsource work to globally distributed and diverse workers [47].

In crowdsourcing platforms, work is "taskified" [5]. Entering receipts into expense reports, curating data to train an Artificial Intelligence (AI) model, translating texts and tagging words and images, are examples of work that can be easily "taskified". These tasks are carried out by an invisible workforce of "humans in the loop" [18]. Through platforms such as AMT, in this volatile and globally distributed crowd of workers, with varying degrees of expertise and backgrounds [47], people compete against each other to find tasks to work on. The logic behind the allocation of these tasks typically operates on a "First-Come, First-Served" (FCFS) basis [19, 59].

Previous research has argued that FCFS is a convenient method of task allocation because of its simplicity and capacity to decrease task completion time [30]. The approach, however, creates a competitive dynamic in which workers are forced to be constantly alert for new tasks to appear, producing a sense of anxiety and frustration in case they cannot obtain the work [18]. Additionally, FCFS disadvantages workers who do not have access to a reliable Internet connection or those who work in time zones different from the requesters. Thus, it can create inequitable work distribution by relying on circumstances that are often beyond workers' control.

Alternatives of work distribution have been proposed in crowdsourcing literature to optimize worker-task matching, maximising the task-requesters' benefits, and improving results (e.g., [10, 23, 31, 58]), yet, besides some empirical studies that examine crowdsourcing issues from the workers' perspective (e.g., [7, 14, 40]), there is a lack of proposals which aim to improve the working conditions and well-being of workers. This article reports the preliminary results of an interactive design approach where workers have been involved throughout a research process that includes a variety of methods and which aims to identify and validate alternative task allocation logics defined and agreed upon by the workers themselves.

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8500-8/21/09.
https://doi.org/10.1145/3479986.3479987



1. Type what you hear
2. Sync Timing
3. Review and complete

Watch the video again and verify that the subtitles are complete and correct.

Edit title and description.

If you want to help animals,	2:31.77
never forget that helping people	2:33.47
to experience delicious vegan food	2:35.41
is a great place to start.	2:37.09

Figure 1: Screenshot of AOD's subtitling platform, captured on 24th November 2018.

Our vision towards more worker-centric crowdsourcing platforms is summarised by the motto: *"the platform belongs to those who work on it*[1]*"*, which aims at empowering workers to define the rules that govern the distribution of value in crowdsourcing platforms.

The remainder of the paper proceeds as follows. Next, we present our case study, followed by a description of the theoretical concepts that frame the work and a review of related works. Section 5 introduces the methods employed in the study. Later, Section 6 describes the results. A general discussion about the implications of the results is provided in Section 7. We close the paper by presenting conclusions in Section 8.

2 CASE STUDY: AMARA ON DEMAND

Amara is a project which sustains an open and collaborative platform for the creation of subtitles [27]. Examples of organisations employing Amara's platform to create subtitles drawing on volunteer engagement include Khan Academy, Scientific American, and the California Academy of Science [4]. More specifically, our focus in this research is placed on the use of Amara's platform for the creation of subtitles as an on-demand and paid service: Amara On Demand (AOD). AOD was launched in 2013 [60] as a result of the success [27] of Amara's platform (Figure 1 shows a screenshot of AOD's subtitling tool). AOD is organised as a non-profit organisation, under the umbrella of the Participatory Culture Foundation. AOD is inspired by cooperative and commoning practices [18], presenting a remarkable contrast when compared with the market-based logic of other crowdsourcing platforms. While AOD grew into its own enterprise within *Amara.org*, it adopted the values of the original volunteer community [3].

Over the past years, AOD moved from a few linguists to more than nine hundred at the time of writing[2]. The work of linguists in AOD is remunerated and they are organised on a per-language direction basis and in which English operates as the master language. For example, if a customer requires a set of videos in German to be subtitled into Spanish, this will involve the groups German->English and English->Spanish. In order to join AOD, linguists are required to submit a resume, two examples of captioned or translated work, and pass an online interview as well as a test. The test is intended to ensure linguists understand AOD's guidelines, maintaining quality and thus, client satisfaction.

[1] We have adapted this motto inspired by Teodoro Flores's phrase "la tierra es para quien la trabaja" ("the land belongs to those who work it"), which captures the revolutionaries' vision for land reform in the context of the Mexican Revolution (1910-1920) [41].

[2] As self-reported by key members of AOD's core team during the interviews.

An essential part of AOD is the core team that facilitates and oversees the whole production process, coordinating and sustaining the infrastructure required for the successful creation of subtitles and captioning. The core team operates as a central node in AOD, although their members are globally distributed. The core team also monitors linguists' compliance with the rules. In AOD, there are explicit rules, practices, and guidelines to govern participation and foster professionalism. For linguists, this means completing tasks by deadlines, not assigning themselves more than one video "at the same time", and adhering to project-specific rules. Linguists are expected to adhere to them in order to receive payment.

3 THEORETICAL FRAMEWORK

Crowdsourcing has many definitions, but can be captured by the idea of an open call for anyone to participate in an online task [6, 12, 24] by contributing information, knowledge, or skills. The 'crowd' refers to the group of people who participate in the crowdsourcing initiative online. The crowd can, in theory, emerge from anyone online or specific subsets of people. Participation is either voluntary (uncompensated) or for money (financially incentivised). An instance of voluntary crowdsourcing can be found in crowdsourced journalism [1] or crowdsourcing in crisis management [52]. In paid crowdsourcing, participants are compensated per task, as in microtasking on digital labour market-places such as Amazon Mechanical Turk (AMT) [53] or based on performance as in innovation challenges [28].

For this research, we frame our case study as part of the aforementioned broader phenomenon of crowdsourcing. More specifically, we draw on Hansson et al.'s [20, 21] categorisation of the different modes of production in crowdsourcing platforms to frame AOD as a case of *human computing*. *Human computing* crowdsourcing platforms, such as Amara and AMT, are those in which "users do micro-tasks that do not require much expertise, such as transcribing audio and video files, translating texts, or tagging maps [...] [, and in which] individual crowd members usually undertake tasks independently of one another, sometimes even competing for work on this market [...]" [21]. In this study, Hansson et al. [21] draw on Marx's [39] theory of alienation to understand the relationships between participants in crowdsourcing and the role of the platforms employed to mediate in the activities. Marx [39] described four types of relationships: (1) between producer and consumer, (2) between the producer and product, (3) the producers' relationship to themselves, and (4) their relationships to other producers. Applying Marx's theory of alienation to crowdsourcing, Hansson et al. [21] developed a typology of alienation that reveals significant differences between the cases studied. Figure 2, adapted from their work, depicts the cases of Amara and AMT. For example, with regards to the relationship between an individual producer with the rest of the producers, the position of Amara being closer towards the inner circle means there are stronger bonds between producers (linguists, in the case of Amara). For the case of AMT, which is in the fourth outer circle, this position represents a lack of bonds between producers. This typology is not to be understood as mutually exclusive: these concepts and different modes sometimes co-exist within the same platforms and processes. However, this typology is "useful as a way to discuss how participation in crowdsourcing is motivated and to develop tools with a better awareness of different types of relationships and how these modes of productions produce different types of knowledge".

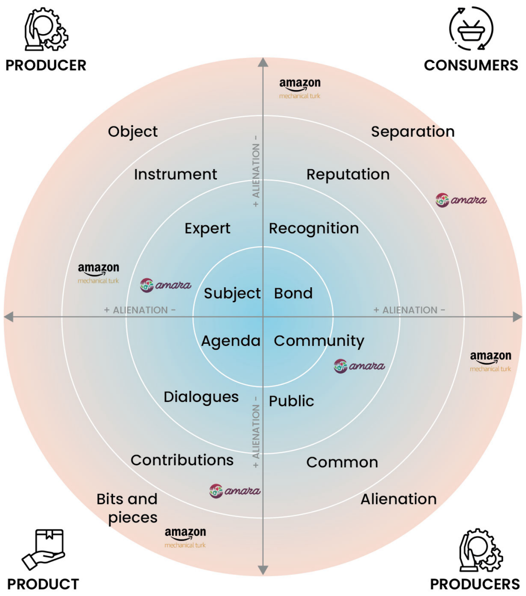


Figure 2: A graphic representation of Hansson et al.'s [21] typology of alienation, according to Marx's [39] four types of relationship. The further from the centre, the higher the degree of alienation. We have adapted Hansson et al.'s [21] Figures 1 and 2 in order to merge the categories and the position of the two key cases (from the 21 studied by them) which we employ to establish comparisons: Amara (our case study) and Amazon Mechanical Turk. See Table 4 on [21] for further details and a summary of the relationships with corresponding modes of productions and the categories employed.

Drawing on Hansson et al.' typology [21] and to further our understanding of how crowdsourcing platforms might support social relationships in these contexts, instead of merely capitalising them [21], we decided to explore the following research question: *can we identify alternative models for the distribution of tasks in crowdsourcing that consider the needs of the workers?*

To this aim, we establish a collaboration with Amara, whose strong cooperative values [18] offer an opportunity to design models of task distribution in which the producer is also the owner of the means of production and the products created are an expression of self-realisation [21]. Since, following the concepts from Hansson et al.' typology [21] depicted in Figure 2, Amara contrasts with platforms such as AMT [18], which understand workers as "instruments" from which "bits and pieces" can be sourced [21].

4 RELATED WORKS

Various alternatives have been proposed in the literature to allocate crowdsourcing tasks [8, 10, 16, 22, 29, 30, 35, 36, 59, 61] as well as to re-think the conditions of crowdsourced labour more generally [17, 32]. Some authors have suggested implementing the power of AI techniques to assign tasks to workers, while other scholars have introduced reputation schemes to delegate tasks. Workers' background, expertise, and social connections have also been considered in approaches presented to improve the assignment of tasks in crowdsourcing platforms.

In Machado et al. [35], the authors suggest using AI planning to help choose the best delegation strategy based on parameters that configure the crowdsourcing environment, such as task duration and workers' skills. A machine learning-based approach that combines supervised learning with reinforcement learning to infer task allocation strategies that best fit the available rewards and workers' reputation is proposed by Cui and colleagues [8]. A content-based recommendation method is introduced in Mao et al. [36] to match crowdsourced development tasks to developers automatically. The system learns from historical activities to favour appropriate workers. Ho et al. [22] present an algorithm to allocate tasks in situations of heterogeneous tasks or a diverse, skilled workforce.

Difallah and colleagues [10] employ information available in the workers' social network profiles, such as their interests, to automatically assign workers to tasks aligned to them. The matching between workers and tasks is based on a taxonomy derived from categories extracted from workers' interests and descriptions of tasks. Likewise, the construction of workers' profiles using historical data of their performance and data extracted from social networks is suggested in Kamel et al. [30]. With these data, the authors propose the development of a machine learning model to recommend relevant tasks to workers based on their built profiles. In a similar way, Zhao et al. [61] discuss a model that considers the relationship between workers to assign tasks in crowdsourcing. The proposal is to use social networking sites to learn about the social connections between workers and therefore allocate them and their friends the same or similar tasks.

The allocation of tasks to groups or teams of workers instead of individuals is explored in [29]. In group-oriented crowdsourcing, members of naturally existing groups of workers cooperate to perform tasks. Jiang et al. introduce, in this article, the concept of contextual crowdsourcing value, which determines the priority of a group of workers being allocated a task. The contextual crowdsourcing value measures the capacity of the group of workers to complete a given task in coordination with other groups that complement the missing skills of the group's members.

Reputation models for task allocation have been studied by [59]. Here, workers' reputation is estimated based on the workers' past performance, considering the quality of previous work and meeting deadlines. Increasing fairness while reducing costs is proposed by Fu and Liu [16] who introduce a task allocation model (F-Aware) to create fairer crowdsourcing workflows. The proposed approach monitors the execution of workflows, adjusting the operation of the allocation algorithm to achieve a fairer distribution of labour among workers.

Our work contributes a novel perspective of task allocation on crowdsourcing platforms. Instead of proposing an approach that targets cost reduction, budget balance, quality assurance, or timely completion optimisation, as in the reviewed literature, we report on alternative models that have the potential to help allocate tasks in a fairer way drawing on co-designing techniques which allow workers themselves to define task allocation models that improve their welfare in the platform. Previous research has also explored collaborations with workers of the platforms to explore alternatives to change the nature of crowdsourcing work. In response to concerns from AMT workers over a lack of employer accountability, Irani et al. [25] developed "Turkopticon." Turkopticon is a platform and browser extension where workers can share experiences about employers, allowing for greater transparency and communication among workers [25]. As part of the tool, Turkopticon reveals workers' views of their task lists with information others have written about employers. AMT workers have also employed generic platforms, such as Reddit, to share advice and experiences of working on AMT [38, 62]. Additionally, researchers in collaboration with AMT workers created a platform called Dynamo to support collective action [51].

However, our study differs from theirs in co-designing directly with AOD workers after establishing a collaboration with the core team that controls and sustains the platform and its code. As discussed in Section 3, AMT and AOD represent different forms of crowdsourcing platforms, showing us how platforms can increase the alienation of workers, but they can also help to reduce it [21]. In this sense, the core team of AOD is willing to experiment with alternative logics that provide more power to the workers themselves and integrate them into the primary platform. In the aforementioned studies [25, 51], on AMT workers co-designing alternative platforms, the platforms developed represent a form of counter-power, rather than control over the main platform. Next, we provide an overview of the methods employed to follow this co-designing approach.

5 METHODS

This study employs a multi-modal qualitative approach that combines data collected from 25 online and face-to-face (F2F) semi-structured interviews, ten months of participant observation, focus groups, and documentary analysis of 55 documents, mainly internal AOD documents provided to linguists and official blog posts from *blog.amara.org*. Table 1 and Table 2 provides an overview of the main characteristics of the participants with whom we conducted semi-structured interviews and organised focus groups. The table includes their gender, main role[3] in AOD, number of years in AOD, location and language groups they belong to (only for linguists), among others.

The collected data were coded following an ethnographic content analysis approach [2], which involved a continuous process of discovery and comparison of key categories emerging from the

[3] This refers to the main tasks carried out by the participant in AOD. For example, as discussed in Section 2 ,whether they are part of the core team. The term DQA refers to Designated Quality Assurer. DQAs are responsible for managing large, active projects where clients often request special instructions. Further details are discussed in Section 6.2.

Table 1: The participants' main characteristics.

Participant ID	Gender	Main role in AOD	Main translation languages	Country of residence	Years in AOD	Related research method
P_1	Male	Linguist	German	Germany	2	Online semi-structured interview
P_2	Male	Linguist and DQA	Marathi and Hindi	India	3	Online semi-structured interview
P_3	Female	Linguist	Greek	United States of America	7	Online semi-structured interview
P_4	Female	Linguist	Russian and English	United States of America	3	Online semi-structured interview
P_5	Male	Linguist and DQA	Greek	Greece	4	Online semi-structured interview
P_6	Male	Linguist	Arabic	Egypt	2	Online semi-structured interview
P_7	Male	Linguist	Dutch and Spanish	Chile	6	Online semi-structured interview
P_8	Female	Linguist and DQA	Hindi and English	India	2	Online semi-structured interview
P_9	Female	Linguist	Simplified Chinese, Traditional Chinese and Malay	Malaysia	2	Online semi-structured interview
P_{10}	Female	Linguist and DQA	English	United States of America	5	Online semi-structured interview
P_{11}	Female	Linguist	French and Romanian	Brazil	1	Online semi-structured interview
P_{12}	Female	Linguist	Greek	Netherlands	3	Online semi-structured interview
P_{13}	Female	Linguist	Russian	Russia	2	Online semi-structured interview
P_{14}	Male	Linguist	Portuguese-Portugal	Portugal	6	Online semi-structured interview
P_{15}	Male	Linguist	Polish	Poland	2	Online semi-structured interview
P_{16}	Male	Linguist	Swedish	Sweden	4	Online semi-structured interview
P_{17}	Female	Accountant (core)	N/A	United States of America	5	F2F semi-structured interview
P_{18}	Female	Project Manager (core)	N/A	United States of America	7	F2F semi-structured interview
P_{19}	Male	Project Leader (core)	N/A	United States of America	8	F2F semi-structured interview
P_{20}	Female	Project Leader (core)	N/A	United States of America	3	F2F semi-structured interview

Continued on Table 2

data. The various analytical tasks were supported by the Computer-Assisted Qualitative Data Analysis Software NVivo 12.

5.1 Participant observation and interviews

Online participant observation was carried out over six months (October 2018 - March 2019) to engage with the day-to-day practices of AOD linguists: from the recruitment and onboarding processes to the execution of regular tasks, such as captioning. In addition, 17 semi-structured interviews (see P_1 - P_{16} in Table 1 and P_{25} in Table 2) were conducted following a purposive sampling [43] intended to gather the diversity of linguists in terms of language group, experience level, and degree of engagement. The data collected provided us with a rich picture of the experiences, needs and vision of the workflow of an AOD linguist. The primary outcomes of this part of the research were the mapping of the workflow of AOD and the identification of an initial set of communitarian needs which

Table 2: The participants' main characteristics (continuation).

Participant ID	Gender	Main role in AOD	Main translation languages	Country of residence	Years in AOD	Related research method
						Continued from Table 1
P_{21}	Non-binary	Developer (core)	N/A	United States of America	2	F2F semi-structured interview
P_{22}	Female	Project Manager (core)	N/A	Brazil	6	Online semi-structured interview
P_{23}	Non-binary	Recruiter (core)	N/A	United States of America	4	Online semi-structured interview
P_{24}	Female	Customer service (core)	N/A	Spain	6	Online semi-structured interview
P_{25}	Female	Linguist and DQA	Portuguese-Brazilian, English and Spanish	Spain	5	F2F semi-structured interview and focus group
P_{26}	Female	Linguist and DQA	Portuguese-Brazilian	Brazil	4	Focus group
P_{27}	Male	Linguist	Portuguese-Brazilian	Brazil	7	Focus group
P_{28}	Male	Linguist and DQA	Portuguese-Brazilian	Brazil	5	Focus group
P_{29}	Female	Linguist and DQA	Portuguese-Brazilian	Brazil	3	Focus group
P_{30}	Female	Linguist	Portuguese-Brazilian	Brazil	5	Focus group

led us to discover several intervention points as potential areas to experiment with the development of worker-centric tools to support crowdsourced labour.

A similar approach was conducted but this time with core members of AOD. It involved four months of online participant observation (April 2019 - July 2019), eight semi-structured interviews (see P_{17} - P_{24} in Table 1 and Table 2) and documentary analysis of materials generated and posted in the official channels of AOD. As well as with the linguists, the semi-structured interviews were conducted following a purposive sampling [43] with key members of AOD's core team considering the diverse roles in AOD, i.e., project managers, developers, members of the finance team, and project leaders, among others. The data analysis carried out here allowed us to further our understanding of the organisational processes of the workflow and the changes experienced in it over time. The aim was to include all of the different perspectives of the actors involved in the platform, to supplement the information gathered from the linguists. More importantly, the analysis of these data led us to select our point of intervention: task allocation. Task allocation is a necessary precursor to working. As a result, task allocation represents a suitable starting point for envisioning more cooperative labour processes.

5.2 Focus Groups

Interviews and participant observations were followed up with an online two-day workshop that included several focus group sessions (organised in June 2020). A call for participation was disseminated through the official AOD channels, including a short survey to show interest in involvement. From all of the linguistic groups in AOD, we chose the Portuguese-Brazilian due to its high degree of organisational complexity. We selected six linguists (see P_{25} - P_{30} in Table 2) according to their different degrees of experience, since we aimed to have a variety of backgrounds. These focus group sessions allowed us, together with the linguistics, to identify alternative models for allocating tasks. The identified models were subsequently validated by the AOD's core team.

5.3 Ethical considerations

The ethical principles described by the European Research Council [13] were followed, as well as the recommendations from the Association of Internet Researchers [37]. Drawing on these guidelines, we constantly reassessed so that the discovery of any new issues resulted in remedial action. These actions include anonymising participants and references to customers in field notes and transcripts, in addition to the use of information sheets and consent forms to participate in the interviews and the focus groups.

6 RESULTS

Next, we describe the series of problems regarding the current logic of task allocation and the main categories surrounding it, which emerged as key from our analysis: (1) first-come, first-served logic, (2) competitiveness, (3) constantly checking for work, and (4) inconsistent workload. Subsequently, we provide an overview of the three alternative models for task allocation identified in this study beyond FCFS.

6.1 Behind the First-Come, First-Served logic

The "First-Come, First-Served" logic embedded in the platform is the main component of task allocation in AOD. This logic creates a competitive dynamic between linguists to assign tasks to themselves. The following quote, from an interview with P_4, depicts the competitive nature associated with FCFS logic:

> "What I realised very fast is that the competition is absolutely awful, the competition is huge, you need to learn how to get the work."

The competitiveness embedded in this logic becomes even more problematic in a global environment, as in the case of AOD. For example, in theory all linguists belonging to the same language group should have the same opportunity to assign themselves a specific task. However, the reality is that some of them might be in time zones that are less convenient concerning the times in which the tasks are usually posted for assignation. Overall, FCFS encompasses a need for workers to check to find more work continuously, an issue identified also in other crowdsourcing projects [18]. For instance, P_4 explained:

> "I learnt how to be fast and not sleep with my computer, but [to] wake up with my computer right next to me. There is also a difference in the time zones, [and] I think we are in the worst position. [...] It is competitive even just to grab it [a task]. I need at least 7 hours of sleep. [...] If you really want to get this work, you need to be next to your computer for hours."

Furthermore, this FCFS logic needs to be understood in an environment in which the workload is typically inconsistent for most of the language groups, as P_{12} explains:

> "Last year, we only had text work like one month, and the rest of the months we had really short videos, like one minute, two minutes, like an advertisement. So that was it: last year it was poor. But the year before, it was a very, very good year. [...] I would like to make a full wage out of Amara, but I don't have the chance. Maybe later on, if the organisation expands. Because it's cool that you can work from wherever you are, on your own timing."

Some linguists also suggested that the competitiveness embedded in the platform's task allocation method undermines the sense of community. P_7, for example, explained it in the following way:

> "In general, I believe we have a very neutral attitude towards each other because... yes, we share the language. But we are also competing to get the jobs [...]. Especially, in the last couple of years, due to the decline in job orders that I have won, I can tell you that it [competition] is growing."

As we introduce in Section 2, one of the key changes implemented by AOD's core was to limit to "one at a time" the number of tasks that linguists can assign themselves simultaneously. This change in the logic was a counter-measure to avoid platform vandalism (e.g. it was found that some participants implemented computer scripts to assign themselves tasks as soon as they were published) and as a first attempt to distribute work more equally. Nevertheless, as we have seen, this has not been sufficient to avoid competition between linguists. FCFS is not the most efficient way in terms of productivity either, as the members of the core explained. P_{17}, a core member of Amara and one of the key workers responsible for managing the overall organisational processes in AOD, explains the need to increase the throughput and reduce the time provided to linguists to fulfil deadlines:

> "If I'm a linguist, the way it works now is: I'll get a caption [...] whether or not it takes me three days to finish the job, I know in advance it's doable in three days, so I can wait until day three to do it. So I can assign myself on day one and wait until day three and do it. From my perspective and my job [as a manager], I see that as a disadvantage for the company. Because, first of all, the client is going to get it later. [...] also, maybe there was another linguist that could've done it on day one. So, in a way, we do [work 'on-demand'] [...], but not in a way in which Uber is on-demand. [...] And we don't have as many people."

Considering all of the problems discussed, we organised a co-designing workshop with AOD linguists that allowed us to incorporate their perspectives into the tools that mediate their day-to-day practices. Next, we discuss the three main models which emerged from this initiative.

6.2 Exploring and identifying alternative models for tasks allocation

We identified three alternative models for task allocation beyond FCFS: (1) round-robin, (2) reputation-based, and (3) content-based.

6.2.1 Round-robin. Round-robin (RR) refers, in computer science, to an algorithm proposed in the context of operating systems [54] to decide how to schedule multiple processes competing simultaneously for CPU time. In RR, computational processes are assigned similar amounts of computing time circularly. It is one of the simplest and most straightforward solutions to avoid *starvation* in-process execution and it is known as one of the fairest scheduling algorithms [55].

Within the context of the focus group, a parallel with RR emerged when discussing the need to split the work equally between linguists, as P_{28} suggested:

> "What I was thinking was a way that all translators could, uh, work on tasks on Amara so that we could split the work equally between translators so that there would be a similar monthly workload for everyone."

Rather than in the form of a "pure RR", the model was discussed as a starting point that could be customised according to the context to find alternatives in which a more balanced assignation of tasks is achieved. A key aspect related to this model was the "pre-assignation of tasks", as depicted by the following excerpt in which participants P_{27} and P_{28} intervene:

> "P_{27}: I was thinking about... about (sic) her idea of pre-assignment. Like, (sic) every day the linguists would come, and we'd have their inbox, uh, the tasks for that day. [...] it could be that on a certain day, uh, we wouldn't have the time needed for that particular task. So it would be necessary to consider, uh, something like the option to accept or not that particular task. [...]

> P_{28}: [...] So the translator could be contacted like, uh, they will be given 24 business hours to respond and take a task [...] And the priority would be for someone who's behind this monthly workload. [...]"

The main advantage discussed by the linguists regarding this model is that it tackles the competitive character discussed in section 6.1, as P_{28} explained:

> "I really liked P_{26}'s idea of the pre-assignment of tasks because this takes away the competitiveness aspect of task allocation. [...] This could also be integrated [...] with a spreadsheet that would rank: this translator has worked on this many minutes this month. So the priority would be for a translator who has not worked that many minutes. So that we could, um, reach a fair amount of work for everyone."

Indeed, we found that some linguists in Amara already use similar informal practices in their day-to-day operating. While most tasks related to translating and captioning are allocated in AOD following a FCFS logic, reviewing (another type of task) escapes this logic in some cases. This alternative allocation occurs particularly for active projects in which clients often request special instructions. These reviewing tasks are carried out by linguists with more experience and selected by Project Managers. This special role in AOD is known as Designated Quality Assurer (DQA[4]). As the quote by P_{28} below depicts, within this specific scope, linguists themselves employ a similar RR logic to balance the workload between themselves:

> "I thought about this because [working as a DQA] I developed a spreadsheet that would sum up all the videos that were available for us to work on, uh, so that we could know which video to allocate to whom like, uh, there's a new video. So if *Emma* was about, I don't know, 20 minutes, uh, shorter than I was, then the video would be given to her. And then the next one would be given to me. And we would find a balance between this workload."

As we shall discuss in Section 7, the challenge of this model lies in identifying the specific parameters to encode in these forms of RR allocation and in providing the linguists with mechanisms that enable them to reach consensus among themselves. Furthermore, the parameters of this model could be combined with those from other models, such as the reputation-based system (presented next). Reflecting on these issues, P_{27} and P_{30} explained:

> "P_{27}: [...] there isn't always a new video to work on. So I think the second part we might have to work on, uh, and I think this could also tie in with my suggestion to split the work equally. So based on the, um, background and the ratings of, uh, translators, they would be pre-assigned to new tasks.[...]
> P_{30}: [...] I think attention to deadline is important, and it doesn't matter if you send like a spreadsheet, tell me how many hours, can you work [referring to the calendar idea]. They put like a thousand hours, and then in the day-by-day, you see that they can't, uh, comply to that."

Next, we discuss the model of allocation by reputation to which the linguists refer.

6.2.2 *Reputation-based.*

Reputation systems have been proposed to build trust among Internet users. They are based on collected and aggregated feedback about users' past feedback and help to foster trustworthy behaviours, assess credibility, and discourage dishonest participation [45]. Usually in crowdsourcing platforms, e-commerce websites, and Q&A forums, feedback on users' actions is instrumented through textual comments, numerical rating scores such as one-to-five scales, and boolean evaluations (e.g., yes/no, like/dislike) [46]. Once built, users' reputation is represented as badges, stars, points, or average scores attached to their screen names [33, 44, 57].

In the context of AOD, this category emerged as a model in which tasks are offered to linguists according to the quality of their previous work, based on feedback received by their peers, and depending on the characteristics of the tasks themselves. During the workshop, this model emerged as a *points system.* P_{27}, for example, proposed the following idea for a model:

> "[...] a system that would be able to rank productivity of linguists [...], something like a points system in which the points would be earned, um, with reference to the volume that was processed before, and also based on the quality of previous work [...]"

The main problem tackled by this model, according to the linguists, is that it would help to increase the level of transparency within AOD. Currently, in AOD, several levels operate according to the linguist's experience. The transition between these levels, however, currently lacks clarity and explicit mechanisms, as several linguists pointed out. The following excerpt, from an interview with P_2, illustrates this:

> " [...] About the transition in levels. Translators should know how to reach next levels from the current levels. What's the criteria for that? [...] [we need] transparency on how to move up or just [some] guidelines."

Indeed, as with the previous case of informal practices in RR, the transition between the different levels already operates, although it does so without explicit parameters, as P_{30} explained:

> " [...] We had knowledge of the previous work that was done by certain linguists that had displayed more quality, more commitment. Uh, they have processed more volume. So [they are promoted] based on all this, but [it was] not, not (sic) quantified, so it was more like a qualitative selection."

The challenge of a reputation-based model, as that proposed by the linguists, is arriving at agreements on what to consider within the system. However, within the focus group, linguists found a preliminary consensus regarding its application in the context of the current *AOD levels system.* The system proposed was based on tiers, as the excerpt below by P_{28} illustrates:

> " [...] building a tier system based on the [amount of] minutes of videos that translators have worked on. So, um, that could be, for example, three tiers: novice,

[4] As discussed in Section 5, DQAs are exclusively responsible for the whole reviewing process for large projects. This contrasts with the usual workflow, in which multiple AOD members review videos within a project following a FCFS logic.

> intermediate, and veteran. So the novice [translators] would get shorter non-technical videos, and veteran translators would be given the opportunity to work on longer technical videos. And the intermediate [level] would be a balance between the two.[...]"

The number of possibilities is such that we concluded that specific sessions would be required to fully explore the parameters of this reputation-based model. However, we identified two key characteristics to be incorporated into the design. Firstly, linguists agreed on not only including the final quality of the translation into the system through reviews made by their peers, but the system should also consider if deadlines were met and how well the changes suggested by the reviewer were implemented. The following excerpt by P_{30}, illustrates this:

> "[...] I just wanted to add that not only the quality of work should be considered in the scoring system, but also the attention to the deadline. In the past, we had a lot of problems with translators that were good, but they were terrible with deadlines. That actually made us lose some clients.[...]"

Secondly, the system should promote inclusiveness. For example, linguists expressed their concerns regarding the possible barriers which a reputation-based model could generate. The following quote, from a comment by P_{28}, illustrates this concern with regards to the barriers for newcomers:

> "[...] My only worry is that we should be careful not to exclude newcomers. Uh, for example, if, if (sic) a new person received the low rating, uh, we need to ensure this doesn't compromise how much work they get, because this could be, this could turn into a vicious, vicious (sic) circle as the ones who need to practice the most would not be given enough work to improve on. Uh, so we need to be careful about that. [...]"

Furthermore, linguists envisioned and proposed ways to tackle such challenges in a reputation-based model. For example, they suggested that linguists' degree of experience should be considered to facilitate the allocation of simpler tasks to newcomers to tackle this type of barrier. Other proposals suggested considering a fixed number of recent tasks carried out:

> "P_{30}: [...] So for the newcomers, we should give them shorter videos and simpler subjects. And for the expert translators, the larger videos, longer videos and high profile projects, and high profile clients. [...]
> P_{28}: [...] I also thought about rating based on a fixed number of tasks. Like the five or ten latest videos would be taken into account in this rating system, so that upon working on new projects, your ranking could also improve like so that you don't get affected by the first videos. [...]"

In sum, a model based on reputation would help to tackle the need to constantly check for work and increase the degree of transparency of promotions within the platform. However, the model poses a myriad of challenges regarding inclusiveness or the generation of different, and perhaps more challenging, forms of competitiveness. As the previous excerpt illustrates, the model could not be purely based on reputation. It could instead be combined with content-based assignation characteristics, which could help tackle some of these challenges. The next section explores precisely this in our third model: content-based allocation.

6.2.3 Content-based. Recommendation systems are the cornerstones of modern online services. In social media, they are used to suggest publications [42], in e-commerce sites to offer products [34], in video-streaming applications to recommend multimedia materials [9], and in crowdsourcing platforms to suggest tasks to workers, as we saw in Section 4. Content-based is one of the most widely used techniques employed in recommendation systems. It focuses on matching the characteristics of the artefact to recommend (e.g., topic of publication, movie genre or task description) with attributes of users based on their profiles and historical data [26].

For our case study, this model emerged as one in which tasks are pre-assigned according to two different types of possible matching logics: (1) either the linguist's skills and/or personal preferences regarding specific areas of knowledge, and (2) the linguist's previous experience in the platform concerning the complexity and/or size of the task.

The former initially emerged from discussions on how to ensure the quality of translations, as the following excerpt by P_{26} depicts:

> "[...] if people work based on their backgrounds, they're much more used to [the] terminology and that, in the end, increases the quality."

The initial ideas revolved around attempts to match the linguist's skills to the content of the video to be subtitled. For example, some of the participants in the focus group possessed a degree in Law. Therefore, some argued that the model should prioritise them to complete tasks involving videos concerning Law. This proposal, however, did not reach a consensus. As other linguists argued, sometimes they prefer to work on contents that are not part of their official background:

> "P_{27}: [...] not working only on what we are already specialised in, but having the chance to learn something new. Sometimes that's even the main motivation: entering a new field, uh, learning a new subject, dealing with a completely different area, different worlds. Like we come from from (sic) Law. And we are working as linguists precisely because we didn't like Law that much (laugh).
> P_{30}: [...] I think not only [the] backgrounds of the translators should be considered, but also [their] interests. You know, people have hobbies, people like some sort of movies and [some] kinds of stuff more than others. So I think we should also be considering not only like, you know, university backgrounds, but also personal interests."

As with the case of the reputation-based models discussed in subsection 6.2.2, we concluded specific focus groups would be required in order to explore the details of this model. Nevertheless, an initial consensus emerged about considering the contents of previous tasks as a possible avenue to explore and implement this type of model:

> "P_{27}: [...] we could define this background and this specialisation of interests based on the most recent work [...] their most recent work in the previous weeks or the previous months [...]
> P_{25}: [...] maybe considering also the size [...] like the five last big videos that the person has worked with would define their, (sic) their interests or their, (sic) [areas of] specialisation."

The session concluded by discussing logics which could, therefore, be potentially merged with those from reputation-based models, in which the complexity and the size of the task would also be considered when carrying out this content-based assignation, as the excerpt below including an intervention by P_{26} depicts:

> "[...] I think that may be one way to do that. Eh, like new people would get more priority on smaller jobs, and then people with more experience get priority with larger jobs. And then people, as they get experience, they start working on longer videos. [...] we know that if a person has had little experience before, then [it] is a bit complicated. So maybe the way would be prioritising smaller videos to newcomers, so everybody would have a chance to improve and to learn."

As with the previous cases, content-based models have the potential to be helpful in overcoming some of the problems for workers derived from the predominant FCFS logic, as those described in Section 6.1. The three identified models presented in this section are, however, to be understood as ideal types [56]. These models are not necessarily a description of reality as such, but valuable concepts when employed as methodological tools to systematise and consider facts that enable us to analyse and intervene within this specific social context. Therefore, identifying these models provides us with a helpful starting point to structure and guide our research, enabling the possibility of establishing comparisons between them and identifying specific characteristics that facilitate merging features of one model with the characteristics of another. Next, we discuss our results concerning the previous literature in the area and provide an overview of future avenues for research with the models mentioned above.

7 DISCUSSION

Conceptualising more egalitarian working conditions in crowdsourcing in collaboration with workers is not unique to this study [17, 32]. Nevertheless, our focus on alternate forms of task allocation *within* a cooperative crowdsourcing organisation — as opposed to platforms like AMT — is. Most large crowdsourcing platforms are less cooperative than AOD, making it challenging to alter aspects of the platform. This study illustrates the potential of co-designing hand-in-hand with workers and the significance of task allocation in dictating the nature of work.

Previous research has criticised algorithmic task allocation — typical in ridesharing platforms like Uber and Lyft — for its opacity and control over workers' profiles, routes, customers, and wages. As reported by Gray and Suri in [18], and confirmed in this research, current task allocation methods used in crowdsourcing platforms, which are primarily governed by an FCFS logic, generate competitive dynamics that result in an unfair and unequal distribution of labour and a sense of frustration among workers. We also found that the inherent competitiveness imposed by FCFS also harms the relationship between workers, negatively impacting the sense of community.

Our research in collaboration with AOD's workers allowed us to define alternative methods to FCFS for task allocation. Through a multi-modal methodology that included interviews, participant observation, focus groups and documentary analysis, we identified together with AOD's workers three models that could allocate tasks considering their needs.

The review of the literature demonstrates that most task allocation approaches are focused on optimising the needs of task requesters. These task-requester-centric solutions are intended to reduce costs [23], maximise matching [35], and increase the quality of results [59]. Fairness in task allocation has been discussed by Fu and Liu [16]. However, their solution focuses on creating fair crowdsourcing workflows, i.e., a logically related series of tasks, by minimising costs. In contrast, our work aims to shift the current FCFS logic of task allocation for models designed by workers to create cooperative and more equitable working conditions.

In this sense, round-robin (RR), one of the identified models, was proposed by the participants to create a more balanced workload. In RR, tasks are pre-assigned to workers in rounds, reducing the competitive dynamics of the current FCFS allocation practice. Although there was consensus among participants that having a model that pre-assigns tasks in rounds could improve their experience in the platform, there are still essential implementation details in question. In the forthcoming focus groups with linguists, we will examine the rules and parameters that will define the operation of the model. For example, how does the availability of workers impact the model? How does the complexity of tasks affect the assignation of labour? How does worker expertise influence the distribution of tasks?

In discussing workers' expertise, the participants expressed that a reputation-based model might be a valuable complement to RR. In this context, workers' reputation would be captured through a system that reflects workers' performance based on historical feedback. Although performance-based reputation systems are already part of crowdsourcing platforms (e.g., Waze, TaskRabbit, AMT), participants see this model as applicable to improve transparency in the promotion mechanisms used in AOD. Here, workers need to build a "career path" by moving through different stages to reach higher levels of responsibility within the organisation. The significant concern with the reputation model, according to participants, is discrimination against newcomers and low-rated workers as well as task concentration by highly-rated workers. Brawley and Pruy have also mentioned the harmful impacts of reputation systems in crowdsourcing platforms in [7]. The participants suggest distributing more straightforward tasks to novices to promote inclusiveness, ensuring that they have assigned tasks when joining the platform. As with the case of RR, there are implementation details to be addressed, and we plan to organise specific focus groups with linguists to explore them. It is still unclear, for example, what aspects of workers' history should be factored into one's reputation. Apart from workers' level and their historical performance, some participants suggested other elements such as attention to past deadlines, caring work [48] (e.g., welcoming and tutoring newcomers), and consideration of colleagues' feedback. Gamification techniques based on artefacts that represent participants' reputation (e.g., badges, points, ranking) [15] were also discussed as potential features to equip reputation-based models.

A task allocation model that includes features of the tasks (e.g., complexity, size, topic, length) and workers' skills, background, and preferences was also seen by the participants as a suitable complement to reputation systems. Apart from complementing the reputation model, this method has been found advantageous to match workers with their expertise and skills resulting in high-quality tasks. Alternatively, some participants indicated that assigning tasks that are not necessarily related to workers' abilities and background might favour learning opportunities for workers. The parametrisation of this content-based model was also left for the next round of focus groups with the participants. It remains unclear how workers' domain of expertise should be included in the model and how to operationalise workers' preferences and interests.

The results presented in this article cannot, however, be generalised and should be understood within the particular case of AOD and similar crowdsourcing platforms. Moreover, due to our qualitative approach, the results cannot be generalised within AOD, considering other groups of linguists may relay a significantly different context, and therefore other models might be more suitable for them. In order to tackle this, we are currently carrying out a longitudinal quantitative analysis of the relationship between users and their activities, drawing on the data already available on AOD's platform. Drawing on this data, we plan to carry out a similar research process with linguists of other language groups (e.g., English-Japanese, English-Italian, English-Spanish).

Furthermore, we want to explore how to develop tools which allow crowdsourcing workers to decide on the models to use in different contexts. Similarly, these tools to support decision-making could be employed to determine collectively how a model could be parameterised. In doing so, we plan to develop collaborative decision-making tools that leverage the affordances [49, 50] of distributed-ledger technologies, such as blockchains, to allow crowdsourcing workers to prioritise parameters within the models which better suit their needs, as well as to decide collectively between the models themselves. In the context of our case study, a blockchain-based solution might enable linguists belonging to different language groups to self-organise in Decentralised Autonomous Organizations (DAOs) and resolve task allocation mechanisms that satisfy their requirements[5]. Given that the results are part of a work-in-progress research endeavour, whether the identified models help to reduce competitiveness and improve working conditions is to be validated.

8 CONCLUSION

This article reports on an endeavour to engage crowdsourcing workers in a multi-modal user-centred research process to identify alternative models of value distribution in crowdsourcing platforms. Three models have emerged as a result of the process, namely (i) round-robin, (ii) content-based, and (iii) reputation-based. Although the proposed models have the potential to improve the workers' experience in crowdsourcing platforms by distributing tasks more fairly, implementation details need to be discussed in subsequent research. Our aim is for this line of research to impact beyond this case study to broader areas of the platform economy. Ultimately, our goal is to foster workers' participation in and ownership of the platforms that mediate their work. Similar platforms owned by cooperatives of workers, distributing tasks and value according to the agreements reached by them, could be envisioned for a variety of areas. An example could be a cooperative of taxi drivers whose organisation is mediated by a platform which they control. The platform could distribute and monitor rides and payments according to the rules defined by the workers within their specific context.

In sum, despite the inherent limitations due to the ongoing nature of this research, the models identified throughout the collaboration with AOD show us that it is possible to envision more cooperative models to distribute work, in which the producers progressively become owners of the means of production and the fruits of their labour expressions of self-realisation [21] and, as a result, *platforms might belong to those who work on them.*

ACKNOWLEDGMENTS

This work was partially supported by the project P2P Models (https://p2pmodels.eu) funded by the European Research Council ERC-2017-STG (grant no.: 759207) and by the project Chain Community funded by the Spanish Ministry of Science, Innovation and Universities (grant no.: RTI2018-096820-A-100). We thank all the members of the AOD community: their help made this research possible. A special mention goes to all the linguists who participated in the focus group and the interviews. We also thank the anonymous reviewers for their suggestions. We are also grateful to all of our friends and colleagues of P2PModels for their helpful comments, guidance and advice. We are particularly thankful to Elena Martínez-Vicente and Paulo Colombo for their help organising the focus group and their overall work in the case study. Finally, we would like to thank Tabitha Whittall for her help in copyediting and proofreading this article.

REFERENCES

[1] Tanja Aitamurto. 2015. Motivation factors in crowdsourced journalism: Social impact, social change, and peer learning. *Social Change, and Peer Learning (October 16, 2015). International Journal of Communication* 9 (2015), 3523–3543.

[2] David Altheide. 1987. Reflections: Ethnographic content analysis. *Qualitative Sociology* 10, 1 (1987), 65–77. https://doi.org/10.1007/BF00988269

[3] Amara.org. [n.d.]. *Amara - Award-winning Subtitle Editor and Enterprise Offerings.* https://amara.org/en/

[4] Amara.org. [n.d.]. Caption, Subtitle and Translate Video. https://amara.org/

[5] Sihem Amer-Yahia and Senjuti Roy. 2016. Toward worker-centric crowdsourcing. *Bulletin of the Technical Committee on Data Engineering* (2016).

[6] Daren C Brabham. 2013. *Crowdsourcing.* Mit Press.

[7] Alice M Brawley and Cynthia LS Pury. 2016. Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior* 54 (2016), 531–546.

[8] Lizhen Cui, Xudong Zhao, Lei Liu, Han Yu, and Yuan Miao. 2017. Complex crowdsourcing task allocation strategies employing supervised and reinforcement learning. *International Journal of Crowd Science* (2017).

[9] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. 2010. The YouTube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems.* 293–296.

[10] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. 2013. Pick-a-crowd: tell me what you like, and i'll tell you what to do. In *Proceedings of the 22nd international conference on World Wide Web.* 367–374.

[5] See [11] for an overview of how organisations have been using blockchain technologies with the aim to decentralise governance.

[11] Youssef El Faqir, Javier Arroyo, and Samer Hassan. 2020. An overview of decentralized autonomous organizations on the blockchain. In *Proceedings of the 16th International Symposium on Open Collaboration*. 1–8.

[12] Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science* 38, 2 (2012), 189–200.

[13] European Research Council. 2018. Ethics Self-Assessment step by step. https://erc.europa.eu/content/ethics-self-assessment-step-step

[14] Rita Faullant, Johann Fueller, and Katja Hutter. 2013. Fair play: perceived fairness in crowdsourcing communities and its behavioral consequences. In *Academy of Management Proceedings*, Vol. 2013. Academy of Management Briarcliff Manor, NY 10510, 15433.

[15] Yuanyue Feng, Hua Jonathan Ye, Ying Yu, Congcong Yang, and Tingru Cui. 2018. Gamification artifacts and crowdsourcing participation: Examining the mediating role of intrinsic motivations. *Computers in Human Behavior* 81 (2018), 124–136.

[16] Donglai Fu and Yanhua Liu. 2021. Fairness of Task Allocation in Crowdsourcing Workflows. *Mathematical Problems in Engineering* 2021 (2021).

[17] Mark Graham and Jamie Woodcock. 2018. Towards a fairer platform economy: Introducing the fairwork foundation. *Alternate Routes* 29 (2018).

[18] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass.* Eamon Dolan Books.

[19] Lei Han, Kevin Roitero, Ujwal Gadiraju, Cristina Sarasua, Alessandro Checco, Eddy Maddalena, and Gianluca Demartini. 2019. All those wasted hours: On task abandonment in crowdsourcing. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 321–329.

[20] Karin Hansson, Tanja Aitamurto, and Thomas Ludwig. [n.d.]. From alienation to relation: Examining the modes of production in crowdsourcing. ([n. d.]). https://doi.org/10.18420/ecscw2017-13

[21] Karin Hansson, Thomas Ludwig, and Tanja Aitamurto. [n.d.]. Capitalizing Relationships: Modes of Participation in Crowdsourcing. ([n. d.]). https://doi.org/10.1007/s10606-018-9341-1

[22] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. 2013. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning*. PMLR, 534–542.

[23] Chien-Ju Ho and Jennifer Vaughan. 2012. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 26.

[24] Jeff Howe. 2008. *Crowdsourcing: How the power of the crowd is driving the future of business.* Random House.

[25] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[26] Folasade Olubusola Isinkaye, YO Folajimi, and Bolande Adefowoke Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal* 16, 3 (2015), 261–273.

[27] Dean Jansen, Aleli Alcala, and Francisco Guzman. [n.d.]. Amara: A Sustainable, Global Solution for Accessibility, Powered by Communities of Volunteers. In *Universal Access in Human-Computer Interaction. Design for All and Accessibility Practice* (Cham, 2014), Constantine Stephanidis and Margherita Antona (Eds.). Springer International Publishing, 401–411.

[28] Lars Bo Jeppesen and Karim R Lakhani. 2010. Marginality and problem-solving effectiveness in broadcast search. *Organization science* 21, 5 (2010), 1016–1033.

[29] Jiuchuan Jiang, Bo An, Yichuan Jiang, Chenyan Zhang, Zhan Bu, and Jie Cao. 2019. Group-oriented task allocation for crowdsourcing in social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* (2019).

[30] Menna Maged Kamel, Alberto Gil-Solla, and Manuel Ramos-Carber. 2020. Tasks Recommendation in Crowdsourcing based on Workers' Implicit Profiles and Performance History. In *Proceedings of the 2020 9th International Conference on Software and Information Engineering (ICSIE)*. 51–55.

[31] David R Karger, Sewoong Oh, and Devavrat Shah. 2014. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research* 62, 1 (2014), 1–24.

[32] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1301–1318.

[33] Robert E Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design.* Mit Press.

[34] Seth Siyuan Li and Elena Karahanna. 2015. Online recommendation systems in a B2C E-commerce context: a review and future directions. *Journal of the Association for Information Systems* 16, 2 (2015), 2.

[35] Leticia Machado, Rafael Prikladnicki, Felipe Meneguzzi, Cleidson RB de Souza, and Erran Carmel. 2016. Task allocation for crowdsourcing using AI planning. In *Proceedings of the 3rd International Workshop on CrowdSourcing in Software Engineering*. 36–40.

[36] Ke Mao, Ye Yang, Qing Wang, Yue Jia, and Mark Harman. 2015. Developer recommendation for crowdsourced software development tasks. In *2015 IEEE Symposium on Service-Oriented System Engineering*. IEEE, 347–356.

[37] Annette Markham and Elizabeth Buchanan. 2012. Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee. http://aoir.org/reports/ethics2.pdf

[38] David Martin, Benjamin V Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 224–235.

[39] Karl Marx. 1959. *Economica* 26, 104 (1959), 379–379. http://www.jstor.org/stable/2550890

[40] Brian McInnis, Dan Cosley, Chaebong Nam, and Gilly Leshed. 2016. Taking a HIT: Designing around rejection, mistrust, risk, and workers' experiences in Amazon Mechanical Turk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2271–2282.

[41] John H McNeely. 1966. Origins of the Zapata revolt in Morelos. *Hispanic American Historical Review* 46, 2 (1966), 153–169.

[42] Alan Menk, Laura Sebastia, and Rebeca Ferreira. 2019. Recommendation systems for tourism based on social networks: A survey. *arXiv preprint arXiv:1903.12099* (2019).

[43] Ted Palys. 2008. Purposive sampling. In *The SAGE Encyclopedia of Qualitative Research Methods*, Lisa M Given (Ed.). Vol. 2. Sage, 697–698.

[44] Maria Papoutsoglou, Georgia M Kapitsaki, and Lefteris Angelis. 2020. Modeling the effect of the badges gamification mechanism on personality traits of Stack Overflow users. *Simulation Modelling Practice and Theory* 105 (2020), 102157.

[45] Paul Resnick, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman. 2000. Reputation systems. *Commun. ACM* 43, 12 (2000), 45–48.

[46] Paul Resnick, Richard Zeckhauser, John Swanson, and Kate Lockwood. 2006. The value of reputation on eBay: A controlled experiment. *Experimental economics* 9, 2 (2006), 79–101.

[47] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

[48] David Rozas, Nigel Gilbert, Paul Hodkinson, and Samer Hassan. 2021. Talk Is Silver, Code Is Gold? Beyond Traditional Notions of Contribution in Peer Production: The Case of Drupal. *Frontiers in Human Dynamics* 3 (2021), 12. https://doi.org/10.3389/fhumd.2021.618207

[49] David Rozas, Antonio Tenorio-Fornés, Silvia Díaz-Molina, and Samer Hassan. 2021. When Ostrom Meets Blockchain: Exploring the Potentials of Blockchain for Commons Governance. *SAGE Open* 11, 1 (2021), 21582440211002526. https://doi.org/10.1177/21582440211002526 arXiv:https://doi.org/10.1177/21582440211002526

[50] David Rozas, Antonio Tenorio-Fornés, and Samer Hassan. 2021. Analysis of the Potentials of Blockchain for the Governance of Global Digital Commons. *Frontiers in Blockchain* 4 (2021), 15. https://doi.org/10.3389/fbloc.2021.577680

[51] Niloufar Salehi, Lilly C Irani, Michael S Bernstein, Ali Alkhatib, Eva Ogbe, and Kristy Milland. 2015. We are dynamo: Overcoming stalling and friction in collective action for crowd workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1621–1630.

[52] Gerald Schimak, Denis Havlik, and Jasmin Pielorz. 2015. Crowdsourcing in crisis and disaster management–challenges and considerations. In *International symposium on environmental software systems*. Springer, 56–70.

[53] Daniel B Shank. 2016. Using crowdsourcing websites for sociological research: The case of Amazon Mechanical Turk. *The American Sociologist* 47, 1 (2016), 47–55.

[54] Avi Silberschatz, Peter Baer Galvin, and Greg Gagne. 1999. *Applied operating system concepts.* John Wiley & Sons, Inc.

[55] Andrew S Tanenbaum and Herbert Bos. 2015. *Modern operating systems.* Pearson.

[56] Max Weber. 1904. Die" Objektivität" sozialwissenschaftlicher und sozialpolitischer Erkenntnis. *Archiv für sozialwissenschaft und sozialpolitik* 19, 1 (1904), 22–87.

[57] Jurgen Willems, Carolin J Waldner, and John C Ronquillo. 2019. Reputation Star Society: Are star ratings consulted as substitute or complementary information? *Decision Support Systems* 124 (2019), 113080.

[58] Xiaoyan Yin, Yanjiao Chen, and Baochun Li. 2017. Task assignment with guaranteed quality for crowdsourcing platforms. In *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*. IEEE, 1–10.

[59] Han Yu, Zhiqi Shen, and Cyril Leung. 2013. Bringing reputation-awareness into crowdsourcing. In *2013 9th International Conference on Information, Communications & Signal Processing*. IEEE, 1–5.

[60] Eve Zelickson. 2019. A Balancing Act: Hybrid Organizations in the Collaborative Economy Empirical Evidence from Amara on Demand.

[61] Bingxu Zhao, Yingjie Wang, Yingshu Li, Yang Gao, and Xiangrong Tong. 2019. Task allocation model based on worker friend relationship for mobile crowdsourcing. *Sensors* 19, 4 (2019), 921.

[62] Kathryn Zyskowski and Kristy Milland. 2018. A crowded future: Working against abstraction on Turker Nation. *Catalyst: Feminism, Theory, Technoscience* 4, 2 (2018), 1–30.

Open data in digital strategies against COVID-19: the case of Belgium

Robert Viseur
University of Mons
robert.viseur@umons.ac.be

ABSTRACT

COVID-19 has highlighted the importance of digital in the fight against the pandemic (control at the border, automated tracing, creation of databases...). In this research, we analyze the Belgian response in terms of open data. First, we examine the open data publication strategy in Belgium (a federal state with a sometimes complex functioning, especially in health), second, we conduct a case study (anatomy of the pandemic in Belgium) in order to better understand the strengths and weaknesses of the main COVID-19 open data repository. And third, we analyze the obstacles to open data publication. Finally, we discuss the Belgian COVID-19 open data strategy in terms of data availability, data relevance and knowledge management. In particular, we show how difficult it is to optimize the latter in order to make the best use of governmental, private and academic open data in a way that has a positive impact on public health policy.

CCS CONCEPTS

• **Collaborative and social computing and tools**;

KEYWORDS

open data, COVID-19, open science, knowledge management, health

ACM Reference Format:
Robert Viseur. 2021. Open data in digital strategies against COVID-19: the case of Belgium. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3479986.3479988

1 INTRODUCTION

In December 2019, an outbreak of pneumonia attributed to a new coronavirus named SARS-CoV-2 started in Wuhan City (capital of Hubei Province, China) [9]. This 2019 coronavirus disease (or COVID-19) led to the strict confinement of Wuhan City by January 23, 2020 and most major cities in Hubei by January 24 [17]. Although locally effective, these confinement measures did not prevent the disease from spreading to mainland China and several dozen other countries, leading the WHO to declare the COVID-19 epidemic a pandemic on March 11, 2020. In Europe, the disease hit Italy hard. Despite the gradual implementation of confinement from March 9, the health care system there quickly reached its limits (due in particular to the limited capacity of intensive care units) while other European countries experienced a similar progression of the virus (with a few days or weeks delay) and implemented confinement measures in turn [30]. In Belgium, confinement was imposed as of March 18, 2020, while a second confinement was implemented as of November 2, 2020 after a relaxation of the measures during the summer vacations. Digital technology was part of the arsenal of responses to fight the spread of the virus [34]: profiling of citizens at the borders, identification of contact cases (tracing), cluster analysis, development of big open data... In this difficult health context, some countries published open data related to the COVID-19 as from the first weeks of the pandemic [23]. Open data was also quickly pointed out as a necessary tool to monitor and anticipate the spread of the virus [35].

The unavailability of open data in Belgium was identified at the beginning of the pandemic as a hindrance to the work of scientists[1]. The publication of open data files is the responsibility of Sciensano, a public institution created on February 25, 2018. The unavailability of certain raw data has led to protests from the academic world, sometimes relayed by the press or by elected officials[2]. Among the academic relays, let's mention Bernard Rentier (see Figure 1) a Belgian virologist, former rector of the University of Liège and active promoter of open access policies: *"Very bad news, the (hopefully temporary) closure of the excellent @Covidata.be, exasperated by the lack of transparency on #Covid_19 information in Belgium. Beware: opacity is the main stimulus for conspiracy theorists"*[3]. However, over time, these open data have proven their usefulness, such as the French website covidtracker.fr, developed by Guillaume Rozier (and linked to the popular service ViteMaDose).

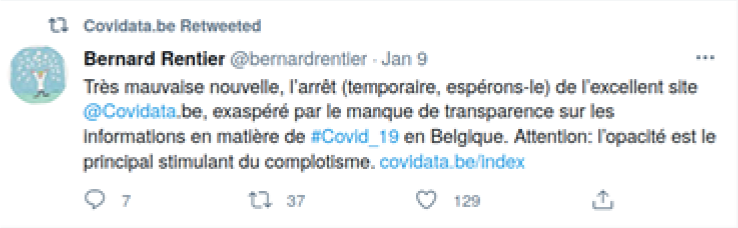
In this article, we propose to draw up an inventory of open data on COVID-19 in Belgium, to identify the uses and the obstacles to reuse, and finally to analyze the available data sets. To do so, we rely on a case study (anatomy of the pandemic in Belgium) based on published data. Our article is divided into four parts. The first part is dedicated to the review of the literature and will be followed by the presentation of the methodology. The third part, divided into three sub-parts, is devoted to the presentation of the results. The fourth part is committed to the discussion of the results.

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479988

[1] Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[2] Cf. http://margauxdere.be/pourquoi-lopen-data-est-important-pour-vaincre-la-covid-19/ or the Twitter account of the Hainaut deputy Catherine Fonck (former doctor)

[3] Cf. https://twitter.com/bernardrentier/status/1347839421899550721



Covidata.be Retweeted

Bernard Rentier @bernardrentier · Jan 9

Très mauvaise nouvelle, l'arrêt (temporaire, espérons-le) de l'excellent site @Covidata.be, exaspéré par le manque de transparence sur les informations en matière de #Covid_19 en Belgique. Attention: l'opacité est le principal stimulant du complotisme. covidata.be/index

Figure 1: Relay of a tweet in support of the Belgian site covidata.be

2 STATE OF THE ART

2.1 Open data

The term "open data" refers to "*information that has been made technically and legally available for reuse*" (19). Data release in the public sector is motivated by transparency (e.g., budget analysis) and the belief that data, like software, are "*public goods*" ([3]; [19]). This practice can also facilitate data sharing between public organizations, foster the emergence of new applications useful to citizens (e.g., public transportation), or improve data quality and enrichment through crowdsourcing initiatives ([3]; [8]; [19]; [20]). Huijboom and Van den Broek [15], for their part, distinguish 3 generic strategies: improving democratic control and representation; strengthening policing and law enforcement; and product and service innovation. The authors show different strategic orientations between states and a more or less pronounced focus in different cases. Open data initiatives in the European public sector are promoted by European Directives such as PSI (2003/98/EC) or INSPIRE (2007/2/EC). Several initiatives are already popular in the UK, the US, France or Brazil (e.g. data.gov.uk, thedatahub.org, data.gov, data.gouv.fr, data-publica.com, www.dados.gov.br...) ([3]; [2]). Nikiforova [23] conducted an early analysis of open data policies related to COVID-19. The results showed that 32 out of 52 national open data portals were publishing such open data while Austria, France, Switzerland, and the US were the most responsive in publishing open data within the first two weeks of the pandemic. These open data can in practice feed into open science initiatives [11]. For example, in the spirit of reproducible research, an epidemiological model is described in a publication accompanied by an open source implementation and open data inputs and results.

2.2 Database

Organizations are required to build databases and operate them as part of a decision-making process [36]. In the case of COVID-19, the organizations involved are those involved in the state health response and must make decisions (e.g., confinement decision and configuration of confinement policy based on hospital congestion status and the extent of virus circulation). In this context, the value of the information will depend on the use made of it and will generally decrease over time. This information is actually a representation of the reality in which we are interested. The representations will be built from a set of signals supposed to translate the states of the real world through an information function [36]. These representations are materialized by the data present in the database. These data are subsequently transformed into information through an interpretation process. This process is based on pre-existing knowledge (in particular the interpretative model used). The use of information leads to results, which are sources of learning and new knowledge, allowing not only the interpretation of data but also their selection during acquisition (filtering).

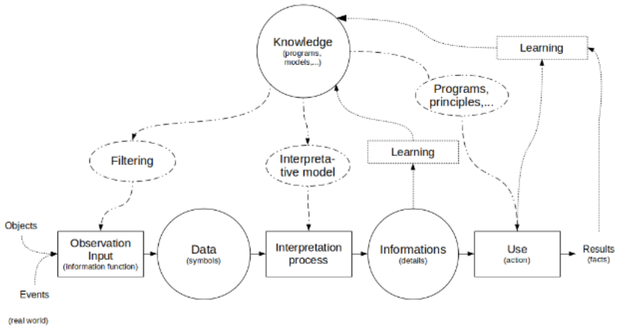


Figure 2: Data valuation process (adapted from Reix et al., 2011)

This process, described by Reix et al. (2011), is shown in Figure 2. It is also described, in a more popular form, as the DIKW pyramid [10]: Data, Information, Knowledge and Wisdom. In this form, data are described as facts, raw materials, that have been accumulated over time "*by people or by machines from observation*"[4]. Information is well-formed data to which meaning has been added. Knowledge involves data, meaning and practice. It is "*a resource for an entity's ability to act effectively*". Wisdom is "*the ability to make optimal use of knowledge to establish and achieve desired goals*". It can be individual (competence) or organizational (capacity). The latter can be related to absorptive capacities [36] or to the dynamic capacities dear to Teece [33].

2.3 Relevance of data

Organizations therefore make decisions based on representations of reality (e.g., management dashboard). Are these representations "*relevant*"? The relevance of representations refers to their use. What is relevant is "*what is appropriate, what is suitable for action*". In other words, data are useful when they lead to individual and collective wisdom. Relevance is therefore "*a quality relative to a user and a context of use*" [36]. Several quality criteria are used to judge the relevance of representations. Reix et al. (2011) cite three main ones:

[4] Translations from references in French are made by the author

completeness, accuracy, and degree of finesse. The quality of completeness refers to the completeness of the representation: all significant changes of state of reality are covered by the representations (first order risk). The accuracy quality refers to the absence of noise: the representations should ideally avoid the second kind of risk related to the taking into account of events caused by random variations due to imperfections of the information function. The degree of finesse concerns the precision of the representation, i.e. its level of detail or its range of variation. In addition, according to Reix et al (2011), there is the punctuality (respect of deadlines), the reliability (confidence in the source), the form (data, drawings, still or animated images...) and the accessibility (search and access methods).

3 METHODOLOGY AND DATA

The Google search engine was used, on the one hand, to identify open data sources related to COVID-19 in Belgium, on the other hand, to identify press articles dealing with open data. In the first case, the queries *open data "covid-19" site:be* and *"open data" "covid-19" site:be* were used. In the second case, the reference news site RTBF Info was targeted using the *"open data" "covid-19" site:rtbf.be* query. Only articles discussing the Belgian open data policy COVID-19 were taken. In particular, RTBF articles simply relaying pandemic figures based on Sciensano open data as part of their information mission were not taken into account. Seven articles developing criticisms on the Belgian open data strategy were thus retained.

The case study was based on the data provided by the governmental reference sites Sciensano and Stabel as well as on the data provided by Google on mobility. The open source software LibreOffice.org Calc (spreadsheet) and R (statistics) were used to process this data.

4 RESULTS: COVID-19 OPEN DATA IN BELGIUM

4.1 Data sources

The unavailability of open data in Belgium was identified at the beginning of the pandemic as an obstacle to the work of scientists[5]. The publication of open data files is the responsibility of Sciensano, a public institution created on February 25, 2018 and the result of the merger between the former "*Centre d'Études et de Recherches Vétérinaires et Agrochimiques*" (CERVA) and the former Scientific Institute of Public Health (ISP).

In practice, the data were published until now in PDF format, which did not allow easy processing in spreadsheet or statistical software. Note that open data are classically published on different portals reflecting the institutional structure of the country. The federal level proposes the portal Data.fgov.be publishing its own data and referencing data from other portals. Other important portals include the portal of the federal statistical institute Statbel, the WalStat portal of the Walloon statistical institute IWEPS and the ODWB portal of the Wallonia-Brussels Federation (FWB).

In practice, the federal portal provides a central view of the characteristics of the datasets made available in open data on its Datasets page. We can see that a large set of formats is supported with a predominance of CSV (textual tabular data), WMS (image format for maps) and JSON (serialization format used by web developers). As for the licenses, they reflect the importance of the referenced datasets: Open Data License Flanders for the Flemish portal, Statbel Open License for Statbel, Etalab Open License for the City of Brussels... As for Sciensano, it publishes its raw data, notably relating to COVID-19[6], on a separate site. The license of the data does not appear clearly on the Epistat site but appears[7] on the FAIR site (Open Data Commons Attribution License). The supported formats are CSV and JSON for the specific data (number of deaths, number of hospitalizations, number of confirmed cases...) and XLS (default format for older versions of Microsoft Excel) for the complete tabular dataset. Although proprietary and supplanted by more recent formats (ODS, XLSX...), the XLS format is still frequently used because it is widely supported as an import or export format for many software programs on the market.

4.2 Reuse of data

These data are important for several reasons. First, they meet communication needs, particularly in the media, in order to explain the characteristics of the pandemic in a more pedagogical way. This is notably the work done by Covidata, an initiative covering a group of researchers proposing analyses and graphics under a CC0 license (CreativeCommons public domain license) accompanied by a set of open source creation scripts published on Github[8] (see Figure 3 for an example of a visualization superimposing the first and second waves with curfew and confinement dates).

Secondly, beyond visualization, they feed into the creation of indicators to better understand the evolution of the pandemic, possibly by combining Sciensano data with data provided by other organizations such as Statbel. We develop an example in section 4.3. There we present a COVID-19 excess mortality index by comparing the proportion of deaths per age group and per region and the respective importance of each age group in the population of each region (see Table 1), which makes it possible to highlight an excess mortality in Wallonia and, above all, Brussels compared to Flanders that cannot be explained by the difference in population density alone.

Thirdly, open data facilitate the development of simulation models that allow a better understanding, and therefore anticipation, of the dynamics of the pandemic at a local level. Nicolas Vandewalle, Professor at the University of Liège, has developed a SEIR model, published on Github[9], which allows to visualize the evolution of the pandemic and to anticipate the saturation of intensive care units, thus providing a tool for decision support (e.g. confinement).

4.3 Case study

The case study proposed here consists of an analysis of the COVID-19 pandemic in Belgium with a specific focus on regional differences. Belgium is indeed a federal state. While health remains an essentially federal competence (e.g. hospitals and social security), on

[5] Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[6] Cf. https://epistat.wiv-isp.be/covid/

[7] Cf. https://fair.healthdata.be/dataset/9c20457b-e66d-4377-ab8e-d48f6dc2b034

[8] Cf. https://github.com/pschaus/covidbe-opendata

[9] Cf..https://github.com/glouppe/covid19be

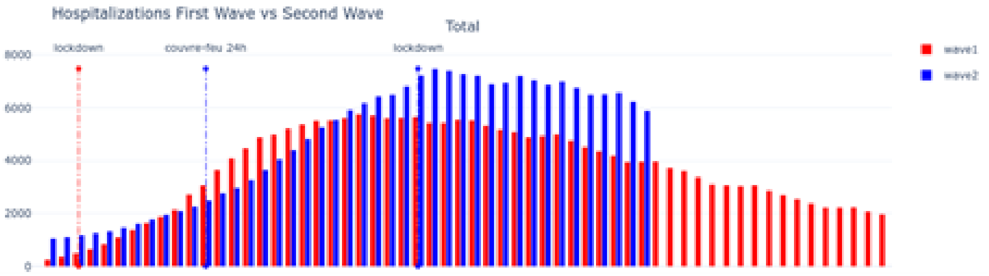
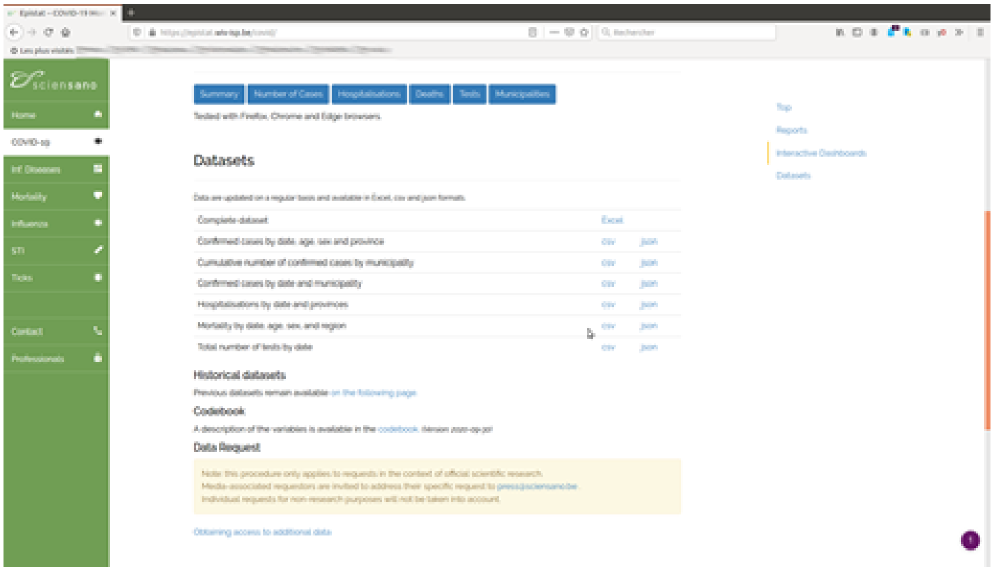


Figure 3: Covidata.be visualization of the number of hospitalizations in open data



Tested with Firefox, Chrome and Edge browsers.

Datasets

Data are updated on a regular basis and available in Excel, csv and json formats.

Complete dataset	Excel	
Confirmed cases by date, age, sex and province	csv	json
Cumulative number of confirmed cases by municipality	csv	json
Confirmed cases by date and municipality	csv	json
Hospitalisations by date and provinces	csv	json
Mortality by date, age, sex, and region	csv	json
Total number of tests by date	csv	json

Historical datasets

Previous datasets remain available on the following page

Codebook

A description of the variables is available in the codebook. (Version 2020-09-30)

Data Request

Note: this procedure only applies to requests in the context of official scientific research.
Media-associated requestors are invited to address their specific request to press@sciensano.be.
Individual requests for non-research purposes will not be taken into account.

Obtaining access to additional data

Figure 4: COVID-19 Sciensano open data portal

Table 1: Mortality by age group in Belgium

Age	Deaths (#)	Deaths (%)	Cumulated
0-24	4	0.03 %	99.85 %
25-44	57	0.42 %	99.82 %
45-64	729	5.38 %	99.40 %
65-74	1626	11.99 %	94.02 %
75-84	3999	29.49 %	82.03 %
85+	7125	52.54 %	52.54 %
NA	21	0.15 %	

the one hand, the overall response to the pandemic depends on other federated entities (e.g. community, for testing in educational institutions, and region for health response in nursing homes), and on the other hand, the regions present distinct social and economic realities.

In practice, Sciensano provides its data in the form of text files in CSV format or workbooks in Microsoft Excel format (see Figure 4). A file containing all the data sets is provided in the latter format. This is the file we work on (dated November 11, 2020). These datasets are provided with documentation on the semantics of the data[10]. The data are generally structured along different dimensions: region, province, city, age (range), gender... The data are of course aggregated, without personal data. The values provided concern the number of deaths, hospitalizations, patients in intensive care and tests.

These files are therefore designed to be easily processed from specific software (e.g. reading CSV files from a Python script) or

[10] Cf. https://epistat.sciensano.be/COVID19BE_codebook.pdf

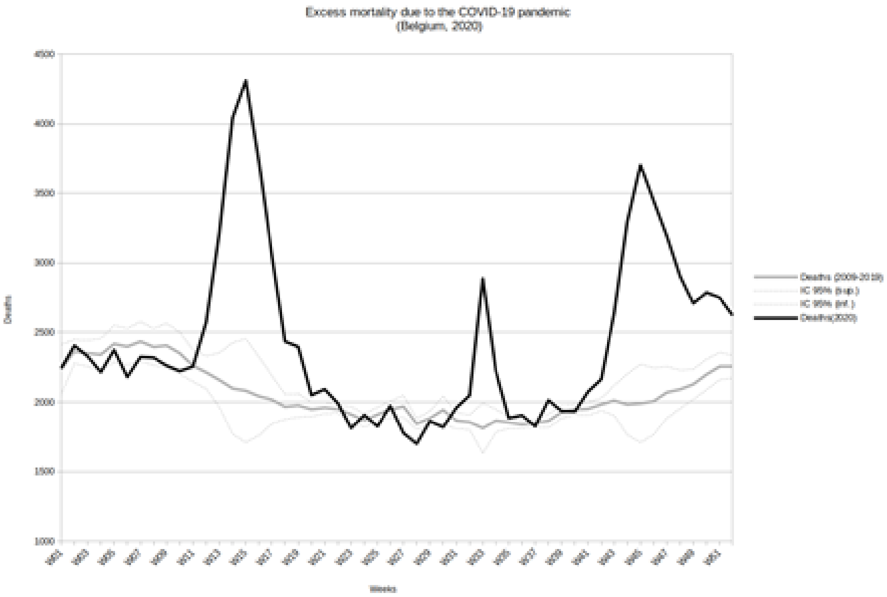


Figure 5: Excess mortality related to the COVID-19 pandemic in Belgium

Table 2: Excess mortality by age group and region

Age	Flanders	Wallonia	Brussels
25-44	-43,45 %	19,42 %	137,26 %
45-64	-45,96 %	38,76 %	149,15 %
65-74	-36,63 %	33,36 %	156,11 %
75-84	-26,48 %	29,64 %	114,69 %
85+	-20,28 %	24,04 %	69,52 %

from standard spreadsheet software using pivot tables or standard functions (NB.SI, SUM.SI...). An open source spreadsheet program such as LibreOffice.org Calc allows for example to easily calculate mortality by age group as well as cumulative mortality starting from the oldest population (see Table 1).

To visualize the existence of excess mortality in the year 2020 during which the pandemic occurred, one has to use the Statbel[11] data provided for the period 2009-2021, which allows to visualize the excess mortality in 2020 during successive waves (cf. Figure 5). These data are also exploited by Sciensano in the Be-MOMO[12] project, for the analysis of excess mortality due to COVID-19 and the validation of the methodology for calculating mortality due to COVID-19 (31) .

The data provided allow the same calculation to be performed by region so as to determine mortality at the national and regional levels (Flanders, Brussels, and Wallonia). However, the comparison is not so simple. Indeed, these regions present distinct realities, notably demographic[13], and we have just seen (Table 1) that the older populations paid the major part of the price in this pandemic. Thanks to the official Statbel website, it is possible to know the population part (%) by age group and by region. From this, it is possible to calculate the excess mortality per age group by dividing the share of COVID-19 deaths and the share of living persons by age group and region (see Table 2). This calculation (i.e. the ratio, for a given province and age group, between the percentage of COVID-19 deaths and the percentage of this sub-population in the general population) shows a high under-mortality in Flanders, given the greater age of the population in the north of the country.

Several factors could explain this excess mortality, such as population density and poverty level. The standard of living, in terms of per capita income, is indeed higher in Flanders[14]. Wallonia, for example, has a high level of unemployment[15] and a lower average per capita GDP (Gross Domestic Product) in its former industrial

[11] Cf. https://statbel.fgov.be/fr/open-data/nombre-de-deces-par-jour-sexe-arrondissement-age

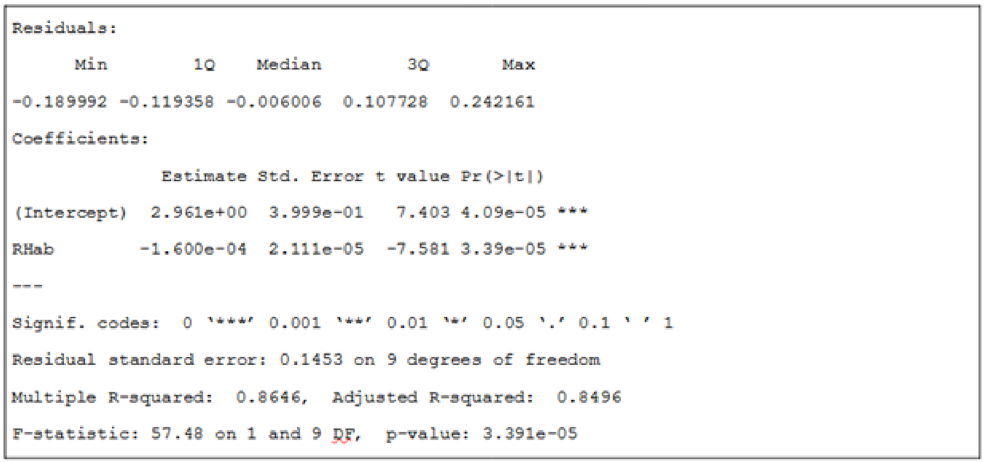
[12] Cf. https://epistat.wiv-isp.be/momo/

[13] Cf. https://statbel.fgov.be/sites/default/files/images/in%20de%20kijker/Chiffrescles_2019_r.pdf, (see page 14 for a breakdown by age group)

[14] Cf. https://www.iweps.be/indicateur-statistique/revenus-menages-habitant/ for a comparison between regions

[15] Cf. https://www.iweps.be/indicateur-statistique/taux-de-chomage-administratif-15-a-64-ans/

Table 3: Linear regression in R



```
Residuals:
      Min        1Q    Median        3Q       Max
-0.189992 -0.119358 -0.006006  0.107728  0.242161
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.961e+00  3.999e-01   7.403 4.09e-05 ***
RHab        -1.600e-04  2.111e-05  -7.581 3.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.1453 on 9 degrees of freedom
Multiple R-squared:  0.8646,  Adjusted R-squared:  0.8496
F-statistic: 57.48 on 1 and 9 DF,  p-value: 3.391e-05
```

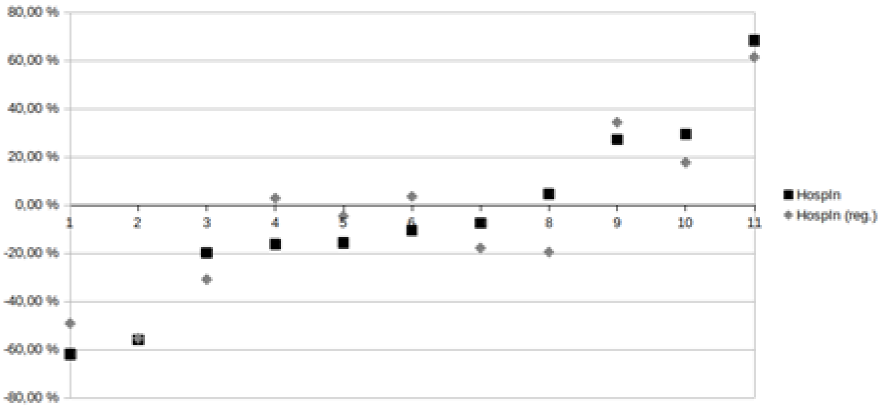


Figure 6: Hospitalization figures (actual vs regression)

area (Mons-Charleroi-Liège axis) [5], while Brussels, although generating a per capita GDP well above the European average [5], is fed by Walloon and Flemish commuters who work but do not reside in Brussels[16], and has a fairly poor population[17] on average, particularly as a result of immigration [25]. Several useful data can be retrieved in this respect: GDP per capita (e.g. IWEPS and Statbel), income per region (Statbel) or specific poverty indicators (e.g. AROPE provided by Statbel). The correlation between per capita income and hospitalization rate, by province, appears to be higher (than that with AROPE or density: -0.930 vs. 0.823 vs. 0.637), so per capita income is used to run a linear regression to predict the hospitalization rate. The latter is used instead of mortality because, at the time of this case (November 2020), mortality by province was not provided, unlike hospitalization figures. The regression model here consists of predicting the excess hospitalization (share of hospitalizations divided by the share of the province's population in the total population) by per capita income. The model provides a satisfactory estimate (see Figure 6): the variable "RHab" is significant and the model has a coefficient of determination R^2 equal to 0.8496 (see Table 3).

The data provided allow for the exploration of other hypotheses (e.g., testing policy; the latter is found to be fairly uniform across the country, for example, regardless of the density of the provinces considered). Other open data can be used, including those provided by private providers such as Google (see Google's "COVID-19 Community Mobility Reports" available for download in CSV format). The latter allow for the analysis of the evolution of mobility related to leisure activities (e.g., increased mobility to the coast during the spring or summer vacations, including the July 21 festivities; see Figure 7) or to work activities (e.g., effectiveness of the confinement and reduction of commuting in Brussels).

4.4 Criticism of the open data strategy

Sciensano's open data policy has been criticized since the beginning of the crisis by prominent French-speaking Belgian personalities such as Bernard Rentier, university pro-rector and promoter of

[16] Cf. https://ibsa.brussels/sites/default/files/publication/documents/FOR_HermReg_2020_12182_F.pdf (page 52 et following)

[17] Cf. https://statbel.fgov.be/fr/themes/menages/revenus-fiscaux

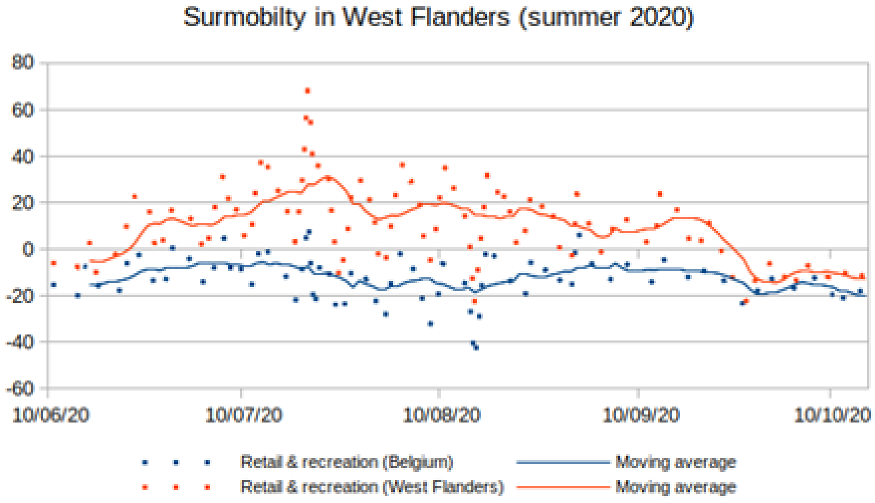


Figure 7: Overmobility at the Belgian coast (summer 2020)

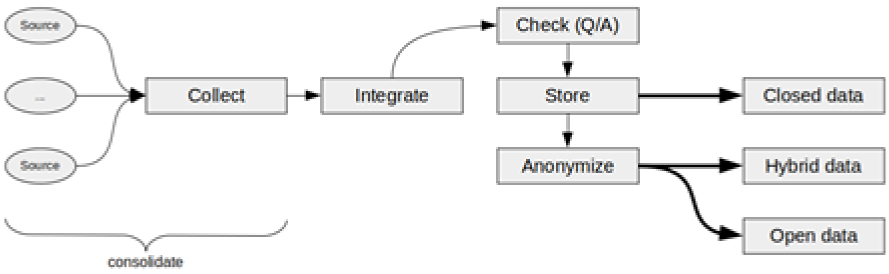


Figure 8: Open data publication process

open science, Pierre Schauss, professor of computer science at the initiative of covidata.be, or Marius Gilbert, epidemiologist: "*The scientific community does not currently have access to the raw data concerning covid-19. This is a major problem that delays us in the answers that can be given to this epidemic. We urgently need to move to open data as Italy and France are doing*"[18]. Basically, Sciensano is mandated to provide data to public authorities.

These early complaints led to the publication of data in open data by the end of March 2020 (the data was then communicated during press conferences or within PDF files that were difficult to process in an automated way). However, tensions with scientists continued until the covidata.be website was temporarily closed in January 2021 in protest against the increasing share of data released outside of open data repositories[19] (a form allows access to sensitive data within three weeks). As Simon Dellicour, an epidemiologist, put it: "*the fact that in Belgium a very small part of the scientific community has access to all epidemiological data is unjustifiable and non-strategic in terms of missed analytical opportunities*" (idem).

As another example, vaccination data were not released until the end of April 2021[20]. In addition to scientists (e.g., studying the effect of returning from vacation), these data are also of interest to journalists for their news and dashboards on the evolution of the pandemic (e.g., explaining the reasons for confinement measures to the population).

Sciensano, on the other hand, points to the workload required to make the collected data available in open data, citing a shortage of 45 people to handle the additional workflow[21]. The difficulty is found at two levels: at the level of the organization itself and at the level of the organization. On the one hand, the collection of data from different sources entails a lot of work to consolidate a network of data collection, to integrate these data and to ensure quality assurance, in particular to validate the published data (cf. Figure 8): "*It would not have been responsible to throw the data away*" (idem).

[18] Cf. https://www.rtbf.be/info/belgique/detail_quand-le-federal-refuse-d-ouvrir-les-donnees-sur-l-epidemie-de-coronavirus-un-probleme-majeur-qui-retarde-les-scientifiques?id=10466339

[19] Cf. https://www.rtbf.be/info/societe/detail_open-data-et-sciensano-manque-de-coherence-entre-le-discours-officiel-et-la-realite?id=10668706

[20] Cf. https://www.rtbf.be/info/societe/detail_campagne-de-vaccination-contre-le-covid-19-le-point-en-chiffres-et-graphiques-ce-dimanche-16-mai?id=10730805

[21] Cf. https://www.rtbf.be/info/belgique/detail_sciensano-notre-premiere-preoccupation-n-etait-pas-de-faire-un-beau-site-internet-en-open-data?id=10509727

Table 4: COVID-19 datasets in open data and by type of source

Government open data(examples)		Private open data(examples)	Scientific open data(examples)
Federal	Sciensano (Epistat, FAIR Healthdata), Statbel[25]	COVID-19 Community Mobility Reports (Google)	Covidata.be (open science: open data + open source)
Region	AVIQ (vaccination statistics in Wallonia[26]		
Community	Open Data Wallonie Bruxelles		
Province	-		
Municipality	-		

On the other hand, data published in open data must be anonymized (e.g., aggregated) and its irreversibility must be guaranteed[22].

5 DISCUSSION

5.1 Characterization of the data sets

The open data identified in this work (see Table 4) are of different natures: either governmental data or private data (made available following the pandemic in the case of Google), or academic data (sometimes in an open science context mobilizing data but also software). The governmental data reflect the Belgian institutional structure but the data specific to the COVID-19 pandemic are fortunately centralized by a unique organization (Sciensano) resulting in frequent redirections to Sciensano from other data sources.

On the licensing side, there are several cases available. In particular, in the case of Sciensano, some of the data is provided on the site under a public domain license while other data is provided upon request: "*Since 31 March 2020, Sciensano has made daily updated data publicly available. Obtaining access to additional data other than these public (open) data is subject to at least: Compliance with the General Data Protection Regulation (GDPR) and the Belgian Law of 30 July 2018 on the protection of individuals with regard to the processing of personal data (including combined datasets consisting partly of personal data and partly of non-personal data); Obtaining authorization of the Information Security Committee, if health data are concerned*".

5.2 Relevance of the representations

Published data raise questions about their relevance (cf. 3.2) in the sense of Reix et al. (2011). The following is a set of potential problems observed following the use of these data or the monitoring carried out on the subject of the use of COVID-19 open data.

Accessibility - Some data are not available in open data (e.g., delays in the availability of vaccination figures in Belgium (and France), added by Sciensano on April 27, 2020. This same observation could be made for the dissemination of wastewater data (cf. Obépine network for example in France). Moreover, some data have been published, but in PDF format, which has handicapped the automated processing on these data.

Accuracy - Test data are affected by the testing techniques that are or are not accounted for (e.g. RT-PCR, antigenic and salivary) in the statistics, which makes comparisons between states or over long periods of time (change of accounting) more complex. In practice, in Belgium, the calculation of the number of cases or the number of tests follows specific rules that Sciensano documents[25]. For example, before March 15, 2020, confirmed cases were in practice, and due to the shortage of tests, possible cases; thereafter, they were cases confirmed by a molecular test (i.e. PCR or rapid antigen test)[26]. Salivary tests, although considered interesting by Sciensano, are therefore not counted[27]. Similarly, the way in which COVID-19 deaths are counted may vary by country and, as in Belgium, may deviate from WHO recommendations [22].

Reliability - Positivity figures should be taken with caution when shortages occur [24]. For example, testing may be prioritized in at-risk or highly symptomatic populations, resulting in higher positivity. Catch-up effects may also occur in the publication of data, leading to upward or downward discontinuities that are not attributable to the observed phenomenon.

Completeness - The effects of confinement measures (e.g. mental health; see [26], on this topic of mental health) are not included in the statistics provided. In addition, data on the use of proximity tracing applications are, as in other countries [34], not public (Belgium).

Fineness - Data are not always provided with the expected granularity (e.g. data by region and not by province in some Sciensano files) but this can be improved over time.

Timeliness - Data are sometimes provided with a delay (e.g. weekend) that disturbs visualization or prediction model results. This problem also exists in Belgium for the monitoring of variants[28].

Punctuality - Sciensano data for Saturday, Sunday and Monday are provided on Tuesday[29].

Form - The processing of open data leads to the creation of new open data as well as to visualizations that facilitate communication around the pandemic (e.g. covidata.be). Graphics can pose different problems (e.g. scale).

[22] Cf. https://www.rtbf.be/info/article/detail_de-sciensano-aux-sites-d-infos-en-passant-par-sydney-comment-les-chiffres-du-covid-arrivent-jusqu-a-vous?id=10717592

[23] Cf. https://statbel.fgov.be/fr/covid-19-donnees-statbel

[24] Cf. https://www.jemevaccine.be/centre-d-informations/un-portail-opendata-vaccination-est-cree-avec-les-chiffres-de-la-vaccination-par-commune-en-wallonie/)

[25] Cf. https://epistat.sciensano.be/COVID19BE_codebook.pdf and https://covid-19.sciensano.be/sites/default/files/Covid19/COVID-19_FAQ_FR_final.pdf

[26] Cf. https://covid-19.sciensano.be/sites/default/files/Covid19/20201012_Advice%20RAG_testing_update%20October_Fr.pdf

[27] Cf. https://covid-19.sciensano.be/sites/default/files/Covid19/20201012_Advice%20RAG_testing_update%20October_Fr.pdf

[28] This information has for example been added to the open data downloadable in France, cf. https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-pour-le-depistage-indicateurs-sur-les-variants/

[29] Cf. https://www.rtbf.be/info/societe/detail_coronavirus-en-belgique-sciensano-ne-publiera-plus-de-bilan-du-coronavirus-les-week-ends-et-le-lundi?id=10525696

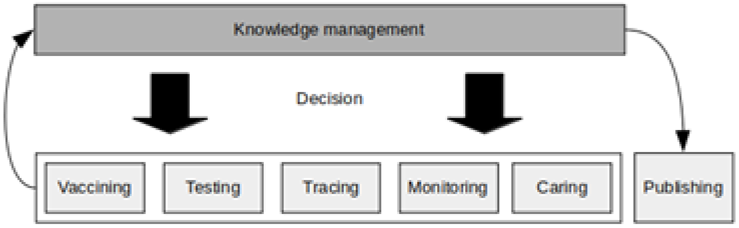


Figure 9: Pandemic response and knowledge management

5.3 Knowledge management

As emphasized by Rowe et al. [29], the effectiveness of the organized response to the COVID-19 pandemic depends on the management of knowledge about the virus to allow for the adequacy and adaptation of measures taken. Thus, the mechanisms of transmission of the virus are the subject of a continuous work of updating of knowledge. In contrast to SARS [13], the importance of presymptomatic persons (persons infected without symptoms during the incubation period but symptomatic beyond that period) in the transmission of SARS-CoV-2, which accounts for more than forty percent of infections, has rapidly emerged ([13]; [32]; [18]). On the contrary, the role of asymptomatic persons (infected persons with no symptoms), about fifteen percent of the patients [14], in the chain of infection may have been downplayed over time ([21]; [4]). As for the modes of transmission of the virus, while there is a general consensus on direct or indirect contact (fomite) and droplets, the possible contribution of aerosols (particles between 5 and 20 μm in size) to airborne transmission was still being debated in March 2021 ([12]; [18]), although this scientific debate does not negate the importance of improving ventilation in enclosed spaces, for example [1].

This evolving nature of knowledge in the face of a new phenomenon implies continuous monitoring and consolidation of knowledge to enable decision-making in the various dimensions of action in the face of the pandemic [34] (see Figure 9). This characteristic also influences the reuse of open data. For example, in the case of models, their execution assumes the availability of relevant data to feed them but also of up-to-date knowledge to configure them, knowledge itself sometimes resulting from a process of progressive data refinement (cf. DIKW model; [10]). Thus, for example, in a model taking into account the diffusion of the virus, the share of contaminations due to presymtomatics must be known and adjusted as soon as the scientific consensus is established or evolves. Similarly, the rate of vaccination, the importance of herd immunity or the importance of immune escape [16] will influence the work of modelers, and presupposes both up-to-date knowledge and the availability of quality raw data.

5.4 Comparison with France

The French situation with regard to the processing of COVID-19 data is also beginning to be documented. Ronai [27] attributes the French delays in publishing complete mortality data to technical and organizational problems. The author distinguishes between a statistically oriented information system, with an administrative channel (INSERM) and a health channel (INSEE), and a real time information system (SI-VIC). The latter, initially deployed to record the victims of the Paris attacks of November 2015, was eventually used for real-time monitoring of COVID-19-related mortality. Initially, it only counted deaths that occurred in hospital and excluded deaths at home or in retirement homes (i.e. 44% of COVID-19 deaths[30]. The latter could finally be counted via a fourth system (Voozanoo) dedicated to EHPADs. The data are therefore partitioned between different systems and must be consolidated before publication. Under pressure from groups of data scientists, such as OpenCOVID19, SantéPubliqueFrance will gradually open up its data sets from March 18, 2020 ([7]; [27]). This will allow several high-profile initiatives to be launched (e.g. CovidTracker). For example, statistics on hospitalizations according to vaccination status were only published by DREES in August 2021 (the breakdown of these data by age group is announced for later), whereas the vaccination campaign started in France on 27 December 2020[31]. They thus required the merging of data from SI-VIC (hospitalization), SI-DEP (screening) and VAC-SI (vaccination)[32]. No matter how much professionals may want to, the production of real-time data sets and dashboards for evidence-based decision-making is not immediate.

6 CONCLUSION

In this exploratory research, we analyzed the open data policy implemented in Belgium. We described the available datasets as well as the applications to which these datasets had led. Based on a case study and a review of press articles, we showed the current limitations of Belgian open data in terms of the relevance of the published data.

This research is currently limited to one country (Belgium). It should be extended to other neighboring countries, which would allow a comparison of publication policies, for example between Belgium and France, which is close to Belgium and is rather among the good students in terms of open data [23], despite the criticisms expressed by the most active reusers[33].

[30] Cf. https://www.lemonde.fr/les-decodeurs/article/2020/12/03/les-residents-d-ehpad-representent-44-des-morts-du-covid-19_6062084_4355770.html)

[31] Cf. https://www.santepubliquefrance.fr/dossiers/coronavirus-covid-19/vaccination-contre-la-covid-19

[32] Cf. https://data.drees.solidarites-sante.gouv.fr/explore/dataset/covid-19-resultats-issus-des-appariements-entre-si-vic-si-dep-et-vac-si/information/?disjunctive.vac_statut

[33] Frenchman Guillaume Rozier, developer of CovidTracker and promoter of open data, thus regularly communicates about the slow pace of data communication (e.g. variant figures). See https://www.lamontagne.fr/paris-75000/actualites/pour-guillaume-rozier-de-covidtracker-l-open-data-permet-de-lutter-contre-la-defiance-et-les-complotistes_13925014/ for an overview.

The understanding of the seemingly more important obstacles to the publication of open data in Belgium would merit further investigation in order to distinguish, and weigh up, cultural, legal, organizational and technical causes. This research could be done through a set of semi-structured interviews in organizations active in health, including IT development structures (e.g. SMALS) and institutions dedicated to public health (e.g. Sciensano).

REFERENCES

[1] Bhagat, R. K., Wykes, M. D., Dalziel, S. B., & Linden, P. F. (2020). Effects of ventilation on the indoor spread of COVID-19. Journal of Fluid Mechanics, 903.

[2] Breitman, K., Salas, P., Casanova, M. A., Saraiva, D., Gama, V., Viterbo Filho, J., & Chaves, M. (2012). Open government data in Brazil. IEEE Intelligent Systems, 27(3), 45-49.

[3] Brugière, A., & Népote, C. (2011). Guide pratique de l'ouverture des données publiques territoriales. FING.

[4] Cao, S., Gan, Y., Wang, C., Bachmann, M., Wei, S., Gong, J., & Lu, Z. (2020). Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. Nature communications, 11(1), 1-7.

[5] Capron, H. (2009). La compétitivité des régions. Reflets et perspectives de la vie économique, 48(1), 115-136.

[6] Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. Nature, 1-8.

[7] Chignard, S. (2021). L'open data de crise : entre mobilisation citoyenne et communication gouvernementale. Enjeux numériques. Les Annales des Mines, n°14, juin 2021, pp. 73-77.

[8] Chignard, S. (2012). Open Data. FYP Editions, France.

[9] Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet infectious diseases, 20(5), 533-534.

[10] Ermine, J. L., Moradi, M., & Brunel, S. (2012). Une chaîne de valeur de la connaissance. Management international/International Management/Gestion Internacional, 16, 29-40.

[11] Fecher, B., & Friesike, S. (2014). Open science: one term, five schools of thought. Opening science, 17-47.

[12] Greenhalgh, T., Jimenez, J. L., Prather, K. A., Tufekci, Z., Fisman, D., & Schooley, R. (2021). Ten scientific reasons in support of airborne transmission of SARS-CoV-2. The Lancet.

[13] He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., & Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. Nature medicine, 26(5), 672-675.

[14] He, J., Guo, Y., Mao, R., & Zhang, J. (2020). Proportion of asymptomatic coronavirus disease 2019: A systematic review and meta-analysis. Journal of medical virology.

[15] Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. European journal of ePractice, 12(1), 4-16.

[16] Kupferschmidt, K. (2021). New mutations raise specter of 'immune escape', Science, 22 Jan 2021 : 329-330.

[17] Lau, H., Khosrawipour, V., Kocbach, P., Mikolajczyk, A., Schubert, J., Bania, J., & Khosrawipour, T. (2020). The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. Journal of travel medicine, 27(3).

[18] Lau, E. H., & Leung, G. M. (2020). Reply to: Is presymptomatic spread a major contributor to COVID-19 transmission?. Nature Medicine, 26(10), 1534-1535.

[19] Lindman, J., & Tammisto, Y. (2011, October). Open Source and Open Data: Business perspectives from the frontline. In IFIP International Conference on Open Source Systems (pp. 330-333). Springer, Berlin, Heidelberg.

[20] Lobre, K. (2012). L'Open Data en 2012 : panorama des risques. AIM 2012, 21-23 mai 2012.

[21] Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E., & Dean, N. E. (2020). Household Transmission of SARS-CoV-2: A Systematic Review and Meta-analysis. JAMA network open, 3(12).

[22] Molenberghs, G., Faes, C., Verbeeck, J., Deboosere, P., Abrams, S., Willem, L., & Hens, N. (2020). Belgian COVID-19 Mortality, Excess Deaths, Number of Deaths per Million, and Infection Fatality Rates (9 March-28 June 2020). medRxiv.

[23] Nikiforova, A. (2020). Timeliness of Open Data in Open Government Data Portals Through Pandemic-related Data: a long data way from the publisher to the user. In 2020 Fourth International Conference on Multimedia Computing, Networking and Applications (MCNA), pp. 131-138. IEEE.

[24] Omori, R., Mizumoto, K., & Chowell, G. (2020). Changes in testing rates could mask the novel coronavirus disease (COVID-19) growth rate. International Journal of Infectious Diseases, 94, 116-118.

[25] Perrin, N., & Van Robaeys, B. (2006). La pauvreté chez les personnes d'origine étrangère chiffrée. Fondation Roi Baudouin.

[26] Pfefferbaum, B., & North, C. S. (2020). Mental health and the Covid-19 pandemic. New England Journal of Medicine, 383(6), 510-512.

[27] Ronai, M. (2021), La construction d'un système d'information épidémiologique, Annales des Mines, n°14, juin 2021.

[28] Rowe, F., Fallery, B., Reix, R., & Kalika, M. (2011). Systèmes d'information et management des organisations. Vuibert.>

[29] Rowe, F., Ngwenyama, O., & Richet, J. L. (2020). Contact-tracing apps and alienation in the age of COVID-19. European Journal of Information Systems, 29(5), 545-562.

[30] Saglietto, A., D'Ascenzo, F., Zoccai, G. B., & De Ferrari, G. M. (2020). COVID-19 in Europe: the Italian lesson. Lancet, 395(10230), 1110-1111.

[31] Sierra, N. B., Bossuyt, N., Braeye, T., Leroy, M., Moyersoen, I., Peeters, I., & Renard, F. (2020). All-cause mortality supports the COVID-19 mortality in Belgium and comparison with major fatal events of the last century. Archives of Public Health, 78(1), 1-8.

[32] Slifka, M. K., & Gao, L. (2020). Is presymptomatic spread a major contributor to COVID-19 transmission?. Nature Medicine, 26(10), 1531-1533.

[33] Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509-533.

[34] Viseur, R. (2021). Stratégies numériques de lutte contre la pandémie de la COVID-19 : quelle place pour les applications de contact tracing ? Congrès InforSID, Dijon (France).

[35] Xu, B., Gutierrez, B., Mekaru, S. *et al.* (2020). Epidemiological data from the COVID-19 outbreak, real-time case information. Sci Data 7, 106.

[36] Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. Academy of Management Review, 27(2), 185-203.

[37] Reix, R., Fallery, B., Kalika, M., & Rowe, F. (2011). Systèmes d'information et management des organisations (6e édition.). Paris: Vuibert.

How makers responded to the PPE shortage during the COVID-19 pandemic: an analysis focused on the Hauts-de-France region

Robert Viseur
University of Mons
robert.viseur@umons.ac.be

Bérengère Fally
University of Mons
berengere.fally@umons.ac.be

Amel Charleux
University of Montpellier
amel.charleux@umontpellier.fr

ABSTRACT

The COVID-19 pandemic led to the confinement of populations in France on the one hand and to shortages of equipment on the other hand (in particular Personal Protective Equipment). The makers therefore mobilized worldwide to produce this medical equipment. In the Hauts-de-France region, a group of makers organized to produce face shields for hospitals, public health and social care institutions and also for retailers. Our analysis of the collaborative messaging room used to coordinate the production of face shields was completed by the interview of active makers. It was based on an original tool-based integrated and hybrid (quantitative/qualitative) methodology. That work enabled us to update the profile of the participants, the intensity of their contribution, the nature of the innovation implemented, the coordination mechanisms, the associated difficulties and the role of technologies in the makers' response.

CCS CONCEPTS

• **Collaborative and social computing systems and tools**; • **Open source model**;

KEYWORDS

COVID-19, makers, fablabs, collaborative innovation, collaborative platform

ACM Reference Format:
Robert Viseur, Bérengère Fally, and Amel Charleux. 2021. How makers responded to the PPE shortage during the COVID-19 pandemic: an analysis focused on the Hauts-de-France region. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3479986.3479989

1 INTRODUCTION

The month of March 2020 was marked for many European countries (Italy, France, Belgium...) by the exponential spread of COVID-19. Due to the lack of anticipation by governments and tensions over supplies of medical equipment, confinement measures were applied, particularly in France and Belgium [8]. The mobilization of the makers was particularly noticeable in their efforts to provide masks and face shields for healthcare workers, but also in their participation in the emergence of more complex open source projects such as respirators.

In the French region of Hauts-de-France, a group of makers mobilized and coordinated through a series of Riot rooms [3], a collaborative messaging client compatible with the open source Matrix protocol, which is a more open alternative to Discord or Slack. The "*3D printed face shields*"[1] room was the subject of an analysis aimed at understanding the type of innovation, the nature of the productions and the coordination mechanisms implemented by these makers.

This research is organized in four sections. In the first section, we present the makers movement and its possible contribution in a crisis situation. In the second section, we present the methodology used. We develop the software used, especially those available in the form of online services, and their integration into a hybrid quantitative and qualitative methodology. In a third section, we present the results. In a fourth section, and before concluding, we discuss the research methodology and the results obtained.

2 MAKERS AND COVID-19

Popularized by [1], the makers movement covers in practice the massive democratization of production tools through, on the one hand, digital manufacturing tools (including 3D printers and laser cutters) and, on the other hand, the development of open source software and hardware, including several popular 3D printing models (e.g. Prusa, Makerbot and Ultimaker).

The maker movement has been accompanied by the creation of third places: techshops, hackerspaces, fablabs... Dedicated to digital manufacturing, the fablabs are supposed to follow specifications defined by the MIT Center of Bits and Atoms (CBA) including the respect of minimal equipment [21]. The term "fablabs" has subsequently been used to designate third places dedicated to digital manufacturing and open to collaborative practices but not necessarily meeting MIT specifications. In addition, fablabs depend on government entities, institutions (public or private) or universities [4]. Therefore the initial objective of sharing can be complemented by more commercial objectives. Fablabs can be seen as "*global*" places because they are both "*embedded in local economic networks*" and subject to global pipelines as soon as, for example, they adhere to the MIT's fablab charter [20].

Fablabs are de facto included in broader structures. [18] has proposed a framework for analyzing the interactions between formal

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479989

[1] The expressions and the verbatim in French were translated by the authors

organizations and informal actors through the concept of "*middleground*". The latter is a meta-platform that links the activities of an underground with those of an upstream. This "*middleground*" is characterized by 4 elements: "places" (physical), "spaces" (cognitive), "events" and "projects". In this model, fablabs can be seen as an element of the "*middleground*" producing digital commons respecting the principles of open source governance [7]. In their analysis of Belgian and French makers COVID-19 projects, [24] note the existence of a structure on three levels articulating large organizations ("*upperground*"), fablabs ("*middleground*") and makers ("*underground*") with a link to global platforms allowing the centralization of knowledge produced according to the principles of open source innovation [17].

The crisis related to the COVID-19 pandemic accentuated the lack of resilience of European states where the risks were underestimated and the response (at the beginning of the pandemic) proved to be ineffective (lack of masks, overloading of hospitals, lack of tests...) imposing a rapid confinement in order to try to drastically reduce the circulation of the virus [8]. In terms of resilience potential, [19] recalls the debate between centralization and decentralization, standardization and local autonomy, control and capacity building, efficiency and responsiveness. The key would thus lie in a planning effort coupled with the creation of processes that stimulate latent resilience by encouraging positive adaptive behavior. In this scheme, belonging to communities, to extended relational networks, is considered in the literature as a resilience factor. Makers form such networks outside of their affiliating organizations. In the past, makers have been able to demonstrate their usefulness in crisis situations. Examples include the manufacture of portable radioactivity sensors (Safecast devices) following the Fukushima disaster in 2011 and open data publication activities for monitoring radioactivity [13]. This was driven by a movement made up of makers, entrepreneurs, companies and above all citizens, it was not limited to the makers alone.

The COVID-19 pandemic highlighted the role of citizens in the production of artefacts (knowledge, accessories, equipment...) useful to help cope with the crisis. It is the case for open data usage [6] and makers productions [5]. The mobilization of the makers within collectives (e.g. "Makers contre le COVID") was the subject of considerable media coverage[2]. As [24] showed, attention on their productions was focused on personal protective equipment (in particular face shields), accessories (e.g. valves and syringe pumps) and respirators (e.g. the Breath4Life project in Belgium and the MakAir project in France). Fablabs often played a central role in these productions, playing a coordinating role and mobilizing their tools to meet the most immediate needs.

Considering this huge investment of makers in the COVID-19 crisis, we analyze here how do they coordinate (especially in confinement conditions), how do they innovate, what were their real outputs and how were they integrated to a more global institutional response?

3 METHODOLOGY AND TOOLS

Our methodology is based on a qualitative approach [15], focusing on the textual content of an online discussion room and the content of semi-structured interviews, supported by automated analysis tools.

3.1 Qualitative analysis

The core of our methodological approach is based on qualitative analysis. Two types of material could be analyzed qualitatively: on the one hand, the exchanges in the "*3D printed face shields*" chat room of the Riot "Hauts-de-France", and, on the other hand, interviews with various protagonists who participated in the design of 3D face shields. The qualitative analysis follows the precepts described by [15]. The latter proposes an implementation of the anchored theorization method. Allowing for a better understanding of the actors, it assumes three types of coding: open, axial and selective. Rejecting both theoretical ignorance and inherited theories, [15] invites a theoretical sensitivity that offers a compromise between an overly strict framing of the research and an insufficiently marked out exploration of the field. The reproducibility that is important in scientific research here concerns the process rather than the outcome. We applied the method, on the one hand, to the textual content of a Riot chat room, and on the other hand, to the content of the interviews.

The observation field of our analysis is the Riot "*3D printed face shields*" messaging room (see 1). This technological tool was used in the Hauts-de-France (France) to coordinate the activity of makers with digital manufacturing machines, in particular 3D printers, capable of producing face shields useful as personal protective equipment.

The second analysis was applied on semi-structured interviews that were conducted with the help of an interview guide and then transcribed. The coding of the interviews was carried out in an iterative way and was accompanied by the progressive feeding of a logbook.

3.2 Technological aspects

This research mobilized two specific tools: on the one hand, Riot (renamed Element) and on the other hand the Cognitive Services of Microsoft Azure.

Riot is a collaborative messaging client based on the open source Matrix protocol. This technology notably enables the creation of chat rooms and can therefore be used within the framework of collaborative projects. It thus provides an alternative to Slack and Discord. These tools were intensively used by the makers during the first confinement (COVID-19) to synchronize their activities in a context of drastically reduced freedom of movement.

Microsoft Azure is a public cloud computing service of IaaS and PaaS types depending on the functionalities used. We are interested here in Cognitive Services, i.e. "*a complete family of artificial intelligence services and cognitive APIs to help you create intelligent applications*", and more specifically in the "Microsoft Speech" and "Language" sections. The first covers speech processing and, in particular, voice recognition, which allows a recording to be transcribed. The second covers the analysis of unstructured text, including in particular the analysis of sentiments and the extraction

[2] Cf. https://www.lemonde.fr/pixels/article/2020/04/23/les-visieres-imprimees-en-3d-une-reponse-des-makers-a-la-crise-sanitaire_6037538_4408996.html for an example.

Table 1: Characteristics of the observation fields

Criteria	Value
Number of participants:	25
Number of items:	1749
Start of activity:	March 28, 2020
End of activity:	May 14, 2020
Duration of activity:	47 days

Table 2: Information related to interviewees

Participants	Function	Duration
Participant n°5	Engineer (school of engineering)	00:41:52
Participant n°1	Engineer (medical research)	02:41:07
Participant n°3	Engineer (school of engineering)	01:54:04
Participant n°10	Entrepreneur (3D printing)	01:08:44
Participant n°2	Engineer (school of engineering)	01:48:17

of named entities. These services benefit from complete documentation, in French and English, accompanied by numerous examples of source codes written in different programming languages (including Python). In this way, the appropriation of these services is greatly facilitated. Pricing is on a per-request basis, but free of charge option is provided for a certain number of monthly transactions, which facilitates testing and encourages use on smaller volumes of data. Overall, the tariff remains below one euro per thousand transactions. To activate the service, users need to associate a payment card with their Microsoft Azure account and then create the resource corresponding to the desired service (e.g. “Text Analysis” and “Speech”) in the Microsoft Azure administration interface. Azure then communicates two access keys and an endpoint to be reused in each of its scripts.

Our analysis therefore focuses on a chat room operated on the Riot platform compatible with the open source Matrix protocol.

Extraction of the content of a Matrix-compatible chat room is possible with the “Matrix-dl”[3] software (available on Github). The latter allows the content of a chat room to be extracted and saved in a weakly structured text file. A pre-processing step (Python script) is therefore necessary in order to segment the day of publication of a message, the time, the sender, the possible recipients and the content of the message. This information can be saved in a “csv” file, not very suitable for very large data sources, but easy to process in Python (“csv” library) and in spreadsheet software such as LibreOffice.org Calc or Microsoft Excel. As the discussion channels studied did not produce more than a few hundred or thousands of messages, this technical solution was retained.

The “csv” file can then be processed in three different ways. The first concerns the measurement of the evolution of the room’s activity, via the number of messages posted per day, and its graphic representation. The second process involves extracting the pseudonym of the contributors and calculating the number of messages posted to the room. The identification of the most active contributors is useful information for the organization of any interview (if fieldwork is planned). The third treatment consists of extracting the relationships between the contributors, through the exchange of messages. These data, extracted in Python, can be exported in “dot” format and then visualized in the Gephi software. The latter can then be used to calculate graph analysis metrics such as Pagerank or betweenness centrality.

The information in the “csv” file can then be enriched by Cognitive Services. Two services are used here: the extraction of named entities and the sentiment analysis. While the first provides a result of poor quality (perhaps due to the approximate syntax of chat rooms), the second makes it possible to identify messages according to their polarity. A distinction must be made here between non-conflicting messages, but carrying positive or negative information, and conflicting messages. The researcher will collect these second messages when conflicts arouse his interest (see [24], for an example). The calculated data are recorded in a second “csv” file.

The second “csv” file is then integrated with the first one in a LibreOffice.org Calc workbook, the concatenation of the date and time of each message providing a key to gather for each message the information useful for the analysis. Conflicting exchanges are easily identified in the spreadsheet by the succession of negative messages in a block of contiguous messages. An average positivity score can also be calculated for each member of the message room.

This analysis using the exchange platform facilitates the identification of important protagonists, due to their prolixity, their involvement in conflicts or their commitment in the general interest projects. They can then be selected for a semi-directive interview in order to gain a more in-depth understanding of the phenomenon being studied. Qualitative approaches (see for example [15]) recommend recording interviews. The speech to text service of Cognitive Services allows the automation of the transcription (count one hour of calculation for two hours of interview). Correcting the automatic transcription requires one hour of work for every 20 minutes of recording and therefore saves about 50% of the time of a manual

[3] Cf. https://gitlab.gnome.org/thiblahute/matrix-dl/

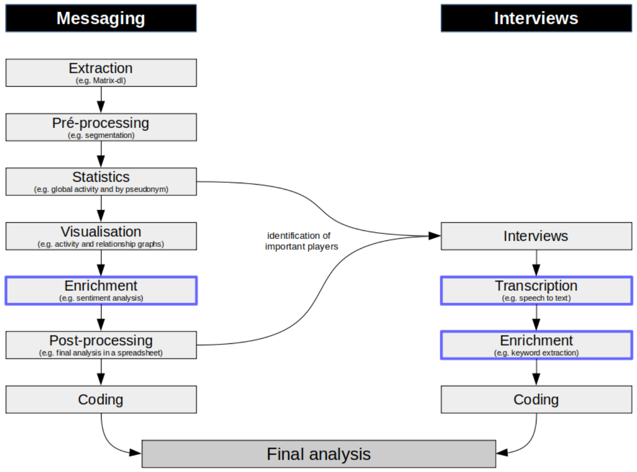


Figure 1: Global methodology and tools

transcription. The document can then be coded [15] after possible automatic extraction of keywords (e.g. named entities). The 2 contains information related to interviewees.

In summary, this analysis therefore combines message extraction with the "Matrix-dl"[4] software (available on Github), their segmentation (Python script), a calculation of activity statistics, sentiment analysis[5] of messages (via Microsoft Azure Cognitive Services), a graph analysis (with the Gephi software) supplemented by a calculation of metrics [12] as well as coding of the 1,749 entries in the collaborative messaging room [15] and lastly the interview with five participants in the discussion room. The segmented data are manipulated and analyzed with LibreOffice.org Calc, which remains valid for discussion channels with no more than a few thousand entries.

3.3 Integrated analysis

The case study relating to the automated analysis of a collaborative messaging room enables us to propose a tool-based analysis methodology (see Figure 1). The rectangles framed in blue highlight the steps that can be automated using Cognitive Services. These, because of their "black box" side, will undoubtedly frustrate researchers who are experienced in text analysis techniques. For the others, these will simplify the prototyping of data processing chains (even if it means replacing them with mastered bricks at a later stage; cf. [23] for some examples of open source reusable components) and will encourage adoption by a wider range of researchers thanks to the lower technical skills required. Indeed, it should be remembered that adoption depends on two dimensions: perceived utility and perceived usability [2].

Collaborative innovation is today widely visible through online exchange spaces (discussion forums, social networks such as Twitter, messaging platforms such as Slack, Discord or Riot...). The analysis of threads can be done manually or automatically. Manually, it can be equipped with qualitative analysis tools such as NVivo and Cassandre. However, these are not dedicated to the analysis of messaging. Given that it's automatic, it relies on collection, backup and analysis solutions, whose complex implementation requires the availability of solid technical skills (see [14] for an example). Some tools are emerging to facilitate the creation of specific processing chains (e.g. R software and its multiple extensions [16]). The tool-based methodology presented here proposes an intermediate path. While programming skills are still required, these remain limited and can be based on abundant documentation with many commented examples of source codes. The result also allows for a control of the processing chain.

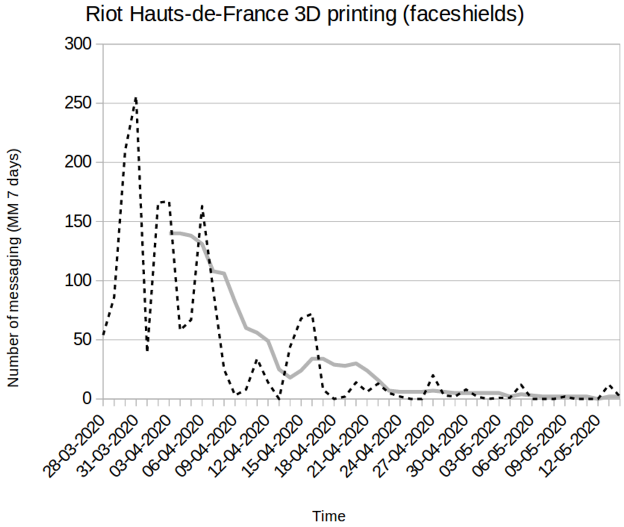


Figure 2: Evolution of members' activity

4 RESULTS FROM RIOT ANALYSIS

The results relate to three distinct aspects: firstly, the activity of the members; secondly, the links between these members (traceable from exchanges); and, thirdly, the nature of the exchanges making it possible in particular to understand, on the one hand, the contributions of this community to the response in times of crisis and, on the other hand, the innovation process implemented.

4.1 Members activity

Analysis of the activity shows that five members produced almost two thirds of the messages posted in this chat room during the period of activity. It thus reflects an application of the law of power of participation [1]. The long tail of less prolific participants is not without interest, however, and includes quality contributions. For example, while the participant n°1, who was responsible for 21.68 % of the messages, played a central role in the design of the face shields, the participant n°8, who was responsible for "only" 3.38 % of the messages and only made more occasional contributions, provided useful expertise in the production and delivery of the face shields (see 3).

Activity was at its peak when the room was first created, before rapidly decreasing (see Figure 2; moving average as a gray solid

[4] Cf. https://gitlab.gnome.org/thiblahute/matrix-dl/
[5] Sentiment analysis makes it easier to identify messages of potential conflict and to identify members who are more involved in such conflicts

Table 3: Members' activity

Participants	Number of messages	Part of messages	Cumulated (%)
Participant n°1	379	21.68 %	21.68 %
Participant n°2	330	18.88 %	40.56 %
Participant n°3	140	8.01 %	48.57 %
Participant n°4	130	7.44 %	56.01 %
Participant n°5	113	6.46 %	62.47 %
Others	657	37.53 %	100 %

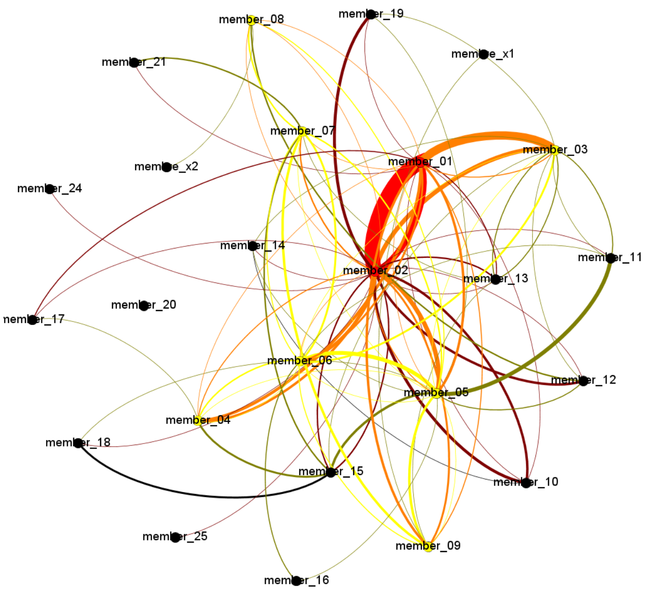


Figure 3: Relationships between members (anonymized)

line). In practice, the design activity shifts to the production and delivery of the face shields (coordinated in another room) once the design has stabilized.

4.2 Links between members

The centrality of a minority of members is reflected in the graphical representation (see Figure 3). The latter makes it possible to visualize a core of the few most active members. The color codes reflect the prolixity of the member (red: more than 250 messages; yellow: less than 250 messages but more than 50; black: others). The thickness of the links shows the importance of the exchanges between members. The participants have more or less specialized contributions. Participant n°2 thus focuses on facilitation, motivation and the dissemination of information. His role as facilitator (and guarantor of the cohesion of the group) is reflected in the highest score for betweenness centrality.

4.3 Relationships, contributions and conflicts

Sentiment analysis allowed us to identify a brief conflict section (about 10 messages) on a classic theme in open source / open hardware / makers environments: respect for licenses and cultural norms specific to the community [10]. In this case: the respect of the NC (non-commercial) clause of the CC-BY-NC license and the profits made (or not) by companies providing face shields created on the basis of validated designs and published under a Creative Commons license. This lively exchange takes as its starting point a trivial question from a service provider (participant n°10) as to the license applicable to a face shield model that he is likely to produce on his machines and the reactions of makers (participants n°1 & n°2) to the margins made by certain service providers: "*I think it's a shame to feel like persona non grata for that*"[6]. This exchange, largely marked by misunderstanding, ends with a reminder of the license applied: "*No, everyone doesn't do what they want. If there are licenses, it's not for nothing. Selling Prusa face shields 10€ is illegal, it's a CC-BY-NC, so it's completely, completely illegal, and we'll have to remember that at the end!*" Moreover, it resulted in a complete disengagement of the targeted member (after a last message that was perfectly neutral).

4.4 Innovation process

The coding of the chat room made it possible to understand the innovation process implemented by this community.

The first stage of this process, which is iterative, is a back and forth process between design and (small-scale) production. The iterations on face shield design are based on feedback from users, manufacturers and health professionals, on comfort of use (head contact, nose size, use of glasses, etc.), sterilization (hygienists; e.g. "*the surface finish is too rough and the hygienist requires something smoother to ensure sterilization by dipping in diluted bleach*") and production (speed, surface finish, strength, post-production... and printing configuration on different machine models). In addition, they also allow for a variety of face shields models to be offered depending on the equipment available, the raw materials and the target audience (e.g. retailers and carers). Designs can be locally innovative and/or based on global designs (e.g. "Dagoma", "Verkstan" and "Prusa" face shields). This step poses a recurring problem of centralizing documentation and designs, a task made complicated by variants and versions. For the "3D printed face shields" Riot room, this was achieved with the OnShape collaborative platform dedicated to computer-aided design, in addition to Etherpad and the room itself (for PDFs). Designs are subject to validation based on feedback from users and reference organizations[7]. Once validated,

[6] The expressions and the verbatims in French have been translated by the authors.
[7] Cf. https://www.onera.fr/sites/default/files/actualites/breves/CASQUE-COVID-19-CNRS-DR14.3-3.pdf for an example cited in the room

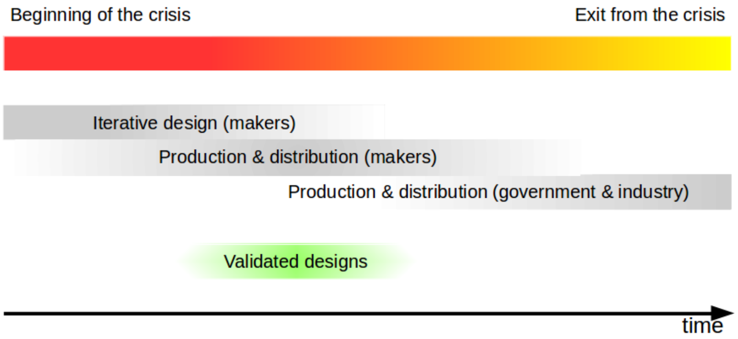

Beginning of the crisis
Exit from the crisis
Iterative design (makers)
Production & distribution (makers)
Production & distribution (government & industry)
Validated designs
time

Figure 4: Temporal representation of the action of the makers

they must then be disseminated within the group and possibly to other platforms working in parallel.

The second stage concerns the move to scale. It involves setting up separate logistics for voluntary and institutional makers (e.g. employees of fablabs and industrial companies). The latter are better able to guarantee a certain level of quality and to coordinate with hospitals. In addition, they have equipment that individual makers do not have (e.g. laser cutters and industrial cutters). This logistics includes in particular the supply of raw materials (in a context of frequent stock shortages), the identification of priority needs (e.g. hospitals), with the implementation (on Google Sheets) of a needs and delivery inventory sheet, and taking into account restrictions on movement (confinement) when delivering to beneficiaries. It is at the level of supplies and withdrawals that collaboration between individual and institutional makers (fablabs, tech shops, industry employees, etc.) comes into play, the latter having travel permits provided by their employers.

The public exhibition is also reflected in the construction of the identity, notably through the name and logo associated with the collective.

This process took place over a month and a half before the crisis situation was resolved, the standards[8] (health, economic, fiscal, etc.) were recalled and the state and industry took over (see Figure 3). The makers thus contribute to the design and selection of one or more dominant designs [22] which can then be mass-produced, on standardized machines, with a low cost price and constant quality, by industry.

5 RESULTS FROM INTERVIEWS

Conducting interviews with people identified, on the one hand, during the content analysis of the "*3D printed face shields*" chat room, and on the other hand, following the graph analysis applied to the exchanges on this same chat room, made it possible to refine the understanding of the coordination of the makers on three levels: the start of the cooperation, the organization of the makers' response and the motivation of the actors.

5.1 Start of cooperation

The creation of the "*Hauts-de-France*" chat rooms came about as a result of the formation of a task force (Polytech, Centrale, CHU Lille) around the problem of the shortage of medical equipment's spare parts, particularly for respirators. Exchanges started with e-mails, then continued on a "*Riot*" chat room (to compensate for the inefficiency of e-mail and to widen the audience), before spontaneously turning to the production of face shields. This evolution can be explained by two factors. On the one hand, the problem of out-of-supply accessories was solved by implementing decontamination procedures and reducing the dependence on disposable products (e.g. RFID chip part that cannot be used after a certain number of uses). On the other hand, the need for personal protective equipment quickly emerged, which motivated the work of the "*3D printed face shields*" chat room. The real objective of the collaboration thus gradually moved away from the initial objective of networking.

> "I think that the Riot was only possible because <participant n°2> was there. That is to say that he very quickly put himself outside the technical discussions, whereas he could have been involved. But he saw very quickly that someone was needed to orchestrate all this. (...) But he took on the role of trying to manage the tensions that might arise or the difficulties, of taking initiatives, of moderating all that, in other words, of creating the chat rooms very quickly. (. . .) In fact, he was there, he was the moderator of the thing, that is to say, he would identify when a discussion was going a bit out of hand. He would advise people to go to such and such a chat room. Or when there was a discussion that was starting to grow, to say wait, instead of polluting this thread, I've created a dedicated chat room for that." (interview: participant n°3)

Eventually, an engineer employed by Centrale took on the role of coordinating the Riot exchange platform. This led him to specialize the rooms, distinguishing three purposes: the production of face shields, the production of accessories and the production of respirators. In the "*3D printed face shields*" room, the collaboration was focused on: face shield design, production optimization and logistics. This organization made it possible to channel energies and avoid the disruption of priority projects (e.g. face shields) by

[8] Cf. https://fabricommuns.org/2020/05/12/realisation-de-visieres-de-protection-nouvelles-normes-et-loi-impactant-les-makers/ for an example cited in the room

peripheral projects whose feasibility and usefulness were more questionable (e.g. respirators), without depriving them of a dedicated collaboration space.

5.2 Organization of the makers' response

The organization can be divided in three steps: the conception of the face shields, their production and then their delivery.

5.2.1 Conception of face shields. The face shield (also called "protective face shield") consists of a headband to which is attached a transparent plastic shield that covers the face. A design of the headband and its fixing to the plastic must be modeled, tested and validated to be optimally printed and assembled. In practice, several designs emerged, which we will refer to as the "Dagoma" face shield, the "CHU" face shield and the "Plastisem" face shield for ease of reference. The general inspiration was given by the face shield proposed by Prusa Research[9] (no feedback to Prusa Research was given). After it was found that this face shield took a long time to print, an adapted model was designed and validated by the nursing staff of the hospital. We will call the face shield "CHU" (also known as "Laurent" face shield in the dedicated lounge, after the designer's first name). For its part, Dagoma chose to produce in maximum quantity (i.e. 10,000 Dagoma face shields in PLA per day at the height of the shortage). The design was therefore conceived so that the face shield's headband structure could be stacked on the 3D printing machines and produced in large quantities. We will call it a "Dagoma" face shield. A third model was then created for plastic injection. The design was based on the "CHU" face shield, which was then modified by Dagoma (3D printer manufacturer) and Plastisem (plastics manufacturer), allowing a mold to be designed quickly and then put into production. We will call this design the "Plastisem" design. 3D printing has since been used profitably not only for production but also for rapid prototyping of molds (allowing very short design times).

> "In fact, there is either an objective to increase production rates, or an objective to satisfy comfort, or an objective to satisfy the technical validation of medical personnel, or an objective to satisfy the ego of one or all of these people. (...) And so in fact there are no good or bad solutions. (...) [Company]'s choice was to say: we are going to give the biggest possible boost in terms of quantities. (...) And so, in fact, we came up with a very thin design because I think that on the last versions, our face shield was 11 grams, so 11 grams to print is necessarily less than 20 grams." (interview: participant n°10)

Other variations, or even completely different designs, were produced. These included designs adapted to specific tooling (e.g. laser cutter) or designs created by makers with expertise in engineering, 3D printing and medical devices (e.g. cartridge mask[10]. These designs could be published on Thingiverse.

> "It's because I was the one who imposed OnShape. So it was hard for me because it wasn't open source. I had a lot of trouble accepting to impose a non-open source thing, but as it was the only distributed CAD tool which meant that you could work in a group without having to install anything. It was good actually. So you see what made it work in the end was the choice of tools and the imposition of a method." (interview: participant n°2)

The "CHU" design was carried out through visible exchanges on the "*3D printed face shields*" room of the Riot "*Hauts-de-France*", completed by frequent evening video exchanges at the initiative of a core group of 5 or 6 people strongly involved in the initiative. The collaborative work was based on OnShape, an online CAD/CAM tool, which is free for non-commercial use, and therefore not open source, unlike Freecad, but which provides a homogeneous collaborative tool. From this tool, an STL file could be extracted for dispatch via other channels (e.g. email). In practice, other designs were developed according to the needs expressed by the audiences consulted (e.g. intensive care staff). The Riot chat rooms were complemented by an Etherpad widget whose content was lost when Riot was renamed Element.

5.2.2 Production and delivery of face shields. Once the designs had been validated and communicated, production could begin. Given the heterogeneity of the 3D printers used, a collaborative effort was required to configure the machines. This was facilitated by the open nature of the open source machines and their use of the gcode language.

> "We had 2-3 users like that who contacted us saying I use such and such a machine but I've never used such and such a material, how could I do it? So I didn't have the machine, but as it's open source, I got a profile. And on this profile I did the slicing and sent them the code that they could install on the machine to make it work." (interview: participant n°3)

Once the face shields were produced, local logistics had to be put in place. In practice, production was centralized at the central pharmacy of the Lille hospital, with a delivery note indicating the model and material used, which then allowed distribution within the hospital (validated model, material compatible with disinfection products) or redirection to other recipients according to the needs expressed. Deliveries were made by a Polytech employee with a travel certificate.

> "I was talking to a reanimator who had worked on the M.U.R. (respirator) project (...) and he was the person who had explained all the constraints to them and so I worked again a little bit with these people,(...) so that I could try to understand the real constraints. And when we understood the real constraints of reanimation, we realized that, in fact, it was useless. There were many things, for example, flow controls which were not ensured even though it was one of their main controls. And the pressure control was really complicated for them to use. So we realized that there were technological challenges for which we should have had a real structured team to do it. This is what MakAir did, for example." (interview: participant n°2)

[9] Cf. https://www.prusa3d.com/covid19/

[10] Cf. https://www.thingiverse.com/thing:4385769)

In addition to face shields, makers have turned to respirators. These are complex medical equipment requiring a minimum of engineering skills and input from specialists in reanimation. The energies on respirator projects, deemed of little use in practice, were channeled into dedicated rooms on the Riot, so as not to disrupt rooms deemed to be of priority (e.g. face shields).

The production of face shields was abruptly stopped (except for some EPHAD needs) after a reminder of the standards in April 2020 by the ANSM. Compliance with these standards would have required certification, which would have been financially costly, and which was not undertaken by any actor involved in 3D printing. This event has had a major effect on the motivation of the makers and clarified their relations as "*underground*" with the "*upperground*" (authorities).

5.2.3 Motivation of the actors. The facilitation of makers communities' platforms necessarily raises questions about the value and the usefulness of the contributions as well as the motivation of the contributors (e.g. do they expect recognition?). Riot's activity was essentially oriented towards the production of face shields, which met a real need in the field. However, some goodwill was directed towards respirators, which posed challenging technical problems, but for which the skills were not available. These contributors were therefore brought together in specific discussion rooms.

> "That is to say that everyone was doing it with bits of string. I mean, nurses ended up with plastic bags instead of scrubs because everything was broken and the day it started again, instead of saying thank you for having provided the interim, we were told to be careful, you're going to have to stop because we could turn against you." (interview: participant n°3)

As for recognition, it came up against the authorities' reaction, which was considered brutal. The end of the crisis left a sometimes bitter taste of non-recognition, or even abuse of the goodwill present at the height of the crisis.

> "It wasn't always a good experience for some people. Because when you go to the central pharmacy to drop off face shields, it's much less rewarding than going directly to the hospital to take a photo with the nurses, which is great for your Facebook." (interview: participant n°1)

As for the small producer-makers, their satisfaction lies sometimes in the recognition of a personal design, which can lead to ego problems, and sometimes in the recognition of a delivered production (e.g. posting a photo on Facebook), which can be countered by optimizing the logistics (centralization). In the case of members employed by a health organization, recognition may have come indirectly from the institution itself (e.g. bonus).

5.2.4 Conflicts between actors. Several fracture lines appeared on the Riot, firstly between the makers in the strict sense of the word and the members from the business world, secondly between the industrialists and the other members (on questions of organization), and thirdly between the "knowers" and the tinkerers.

> "there was a lot of misunderstanding. That's why we made a blog post, so you'll find on our site if you look a bit why we sell our face shields and why it's at cost price. There was a lot of misunderstanding, of "ah bah they make money on the back of the disease"." (interview: participant n°10)

In the first case, the conflict centered on questions of money, in particular the sale of face shields at cost price by a 3D printing company, where the makers gave their production, but where, on the other hand, speculation was noted among certain sellers. The animosity sometimes developed against 3D printing companies led the company to publish a press release which was then taken up whenever the controversy flared up in a discussion forum.

In the second case, the conflict, though less visible in public chat rooms, was about organization and in particular planning issues, with the logic of the manufacturers clearly opposed to that of the 3D printing machine users.

> "For example, on the respirators, we didn't typically have the automation and electronics specialists. We tried to approach them and get them to come in and we didn't get anyone. (...) I think that's also why it didn't follow up too well, whereas in fact we could have found someone who had this knowledge, we could have gone further. In fact, we very quickly came up against technical aspects" (interview: participant n°1) "the people who succeeded were really a team of engineers who made the MakAir, you see, you had people who really worked with engineering methods." (interview: participant n°1)

In the third case, conflicts arise between tinkerers acting by trial and error, sometimes making bad technical choices, with little support for prescriptions from experts. More complex objects such as respirators thus reveal the limitations of the "*make-to-learn principle*" [7] and the need of "*engineer mindset*" to build on solid technological foundation. The Riot messaging platform's exchanges proved to be less conflictual than Facebook-type social networks, which could be explained by the over-representation of a professional audience.

6 DISCUSSION

We discuss the profile of makers observed in the chat room and the future evolution of the innovation platform.

6.1 Types of makers

The expression "*makers' response*", which is widely used to describe the phenomenon under study, gives an impression of homogeneity in each contributing maker's role and profile, whereas the makers are diverse in nature and are divided into co-operative groups. We thus observe, on the one hand, isolated makers with limited resources (equipment, raw materials, etc.) and, on the other hand, sponsored makers, acting with the authorization (more or less formal) of their employer (fablabs, universities, small companies, big manufacturers, etc.) and with greater resources (more expensive equipment, stocks of raw materials, specific raw materials, etc.). What was highlighted by [25] is mainly the organization of institutional makers generally attached to a "*middleground*" in the sense of Simon (2009), in this case a fablab. However, they have collaborated with the makers in the classical sense of the term, on the one hand, through key people, present in several communities, sometimes

Table 4: Profiles of makers

Type	Affiliation	Goal
Pure maker	Hobbyist	Acknowledgment and/or fun (hobby)
Institutional maker	Public-sector	Suitability
Entrepreneurial maker	PME/PMI	Responsiveness
Industrial maker	GE	Efficiency

acting as "*knowledge brokers*", and on the other hand, through the urban logistics that have been put in place. The study of logistics allows us to distinguish a double flow with, on the one hand, centralized delivery rounds and, on the other, peer-to-peer delivery rounds. The latter were carried out by makers transferring parts from one place to another, depending on the traffic constraints imposed by the confinement, while the centralized tour also supplied certain makers with raw materials in exchange for production. We therefore have a coexistence between different more or less articulated networks federating institutional makers (Riot "*Hauts-de-France*") and makers in the strict sense ("Visières solidaires", "Makers against COVID"...), the latter generally coordinating themselves via social networks (e.g. Facebook).

The logistics of production and delivery of face shields therefore tends to be fragmented between, firstly, isolated makers, secondly, institutional makers and, thirdly, manufacturers. The exchanges within the collaborative messaging system also reflect the cultural differences between these different classes of players. On the one hand, the issue of licensing and production pricing emerges as an important point of attention from the contributors, while industry representatives tend, once the designs have been validated, to be focused on the optimization of the production and the delivery capacities (e.g. "*If you allow me, and without offending anyone, because you are all here to help, and that's great: You have to focus on production, you have to know what your production capacity is, as of today, and your daily delivery capacity.*"). In the end, this observation of the Riot chat rooms allowed us to distinguish 4 types of makers (cf. 4) with a dominance of institutional makers, contrary to what could be found on the Facebook groups more oriented towards makers in the strict sense, leading to sporadic smooth conflicts due to differences in motivations and goals.

6.2 Co-evolution of innovation platforms

Different makers' platforms are evolving here in parallel. Information can circulate between platforms following active members registered on several ones. In the same territory, platforms can operate in parallel, with little interaction, serving different target audiences. At the level of the Riot platform studied, success is based on different factors. On the one hand, there is the collaborative activity of iterative design, sometimes fed by designs proposed by high-visibility organizations. Ideally, it requires an efficient tool for centralizing designs, which has been lacking here (other makers in Belgium and France used Github or Gitlab, for example, to centralize the final designs ; [25]). On the other hand, the social usefulness of the productions presupposes a concerted work at the level of the local ecosystem, which implies frequent interactions with the beneficiaries (needs, constraints, feedback...), the pooling of complementary individual expertise (search for common solutions) and the channeling of the contributors' energy (design, production...).

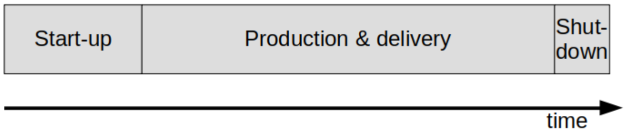


Figure 5: Resilience platform evolution

The platform dedicated to the production of face shields quickly withered away after the first confinement. After a start-up phase during which the individuals working together define the objectives and specify the organization, the platform enables production to be set up before an abrupt halt following a reminder of the PPE standards (cf. Figure 5). In the case of the face shields, although the "*upperground*" helped to stimulate the initiatives of the institutional makers, it did not collaborate formally with them. In practice, coordination with the CHU ("*upperground*") often took place informally and on the initiative of the field staff (informants, doctors, etc.). Contacts with officials generated reassuring speeches (e.g. at the Lille CHU), in contrast to feedback from the field, or very late (e.g. from the ARS). Therefore, if the initial impulse comes from the "*upperground*", the resilience allowed by this maker organization relies largely on the practices and machines available within the "*middleground*" as well as on the informal relationships existing between the "*upperground*" and the "*middleground*".

However, the spontaneous platform did not remain without descendants. Its social usefulness was indeed recognized locally, which made it possible to initiate the setting up of a 3D printing platform for health bringing together the same partners (CHU, Polytech and Centrale, Lille). The resilience platform is thus succeeded by an innovation platform, of which it was in a way the prototype. Moreover we thus believe we can distinguish between temporary production platforms (face shields) and sustainable innovation platforms (respirators). While the former offered a form of resilience in a crisis situation and benefited from a form of tolerance from the "*upperground*" before the latter took control of supplies, the latter led to more formal articulations justified by the complexity of the devices studied (e.g. respirators). This is for example the case of the MakAir respirator project which survived the first confinement and even led to deliveries to India in May 2021[11]. The future will show how much room is left for the "*underground*" in these new platforms

[11] Cf. https://www.ouest-france.fr/pays-de-la-loire/nantes-44000/entretien-une-centaine-de-makair-respirateurs-artificiels-made-in-nantes-produits-pour-l-inde-d3094990-b97a-11eb-a992-89f0a8dfc0f7

that are supposed to bring about active resilience [19], and whether their organization, their animation, develops in coherence with the theoretical framework provided by Simon (2009).

7 CONCLUSION

In this research, we observed and analyzed a community located in Hauts-de-France region, gathered in a Riot collaborative room and mobilized in to produce face shields during the first pandemic. This research is based on a double qualitative approach. The first is based on the exchanges in an online discussion space. The second is based on a set of interviews conducted with selected participants in this discussion space. Five people were identified for their role, commitment, expertise and/or involvement in conflicts. It leads to two contributions. On the one hand, we were able to propose a research methodology combining the quantitative analysis of online communities with the qualitative analysis of messages and the conduct of interviews. On the other hand, we were able to better understand the coordination, contributions and the profile of the mobilized makers.

The main limitation of this research is its localized nature. The analyzed discussion space concerns the Hauts-de-France and is strongly focused on the Lille region. Moreover, created and above all fed by institutional makers, it does not allow us to deeply analyze the organization of pure (isolated) makers in the face of the pandemic.

ACKNOWLEDGMENTS

This research was supported by the European Regional Development Fund as part of the FabricAr3v cross-border project.

REFERENCES

[1] Anderson, C. (2007). La longue traîne. Village mondial.

[2] Bobillier-Chaumon, M. E., & Dubois, M. (2009). L'adoption des technologies en situation professionnelle: quelles articulations possibles entre acceptabilité et acceptation?. Le travail humain, 72(4), 355-382.

[3] Cabanel, C. (2020), Covid-19 : à Lille et en région Hauts-de-France, des makers en première ligne, Makery, 7 juillet 2020 ; en ligne : https://www.makery.info/2020/07/07/a-lille-et-en-region-hauts-de-france-des-makers-en-premiere-ligne/ (consulté le 01/03/2021).

[4] Capdevila, I. (2015). Les différentes approches entrepreneuriales dans les espaces ouverts d'innovation. Innovations, 48(3), pp. 87-105.

[5] Chalet, L., Dutilleul, M., Fages, V., Gayoso, E. (2021). Des visières à haut débit : un regard sociologique sur la mobilisation des makers face à la crise sanitaire. Annales des Mines, n°14, juin 2021.

[6] Chignard, S. (2021). L'open data de crise : entre mobilisation citoyenne et communication gouvernementale. Annales des Mines, n°14, juin 2021.

[7] Cohendet, Patrick, Grandadam, David and Suire, Raphaël (2021). Reconsidering the dynamics of local knowledge creation: Middlegrounds and local innovation commons in the case of FabLabs. Zeitschrift für Wirtschaftsgeographie, vol. 65, no. 1, 2021, pp. 1-11.

[8] Desson, Z., Weller, E., McMeekin, P., & Ammi, M. (2020). An analysis of the policy responses to the COVID-19 pandemic in France, Belgium, and Canada. Health Policy and Technology, vol. 9, pp. 430-446.

[9] Dubois, M., Bobillier-Chaumon, M.-E. (2009). L'acceptabilité des technologies : bilans et nouvelles perspectives, Le travail humain, 2009/4 (Vol. 72), pp. 305-310.

[10] Fauchart, E., Rayna, T., Striukova, L.: Is selling caring? Norms regulating commercialisation and sharing behaviour with the open hardware RepRap. Proceedings of the "XXVIème Conférence Internationale de Management Stratégique" (AIMS), Lyon (France) (2017).

[11] Grandadam, D., Cohendet, P., & Simon, L. (2013). Places, spaces and the dynamics of creativity: The video game industry in Montreal. Regional studies, 47(10), 1701-1714.

[12] Hansen, D., Shneiderman, B., & Smith, M.A. (2010). Analyzing social media networks with NodeXL: Insights from a connected world. Morgan Kaufmann.

[13] Kera, D. (2015). Open source hardware (OSHW) for open science in the global south: geek diplomacy? Open Science, pp. 133-156.

[14] Leclercq, E. & & Savonnet, M. (2018). Modèle tensoriel pour l'entreposage et l'analyse des données des réseaux sociaux. In INFormatique des ORganisation et des Systèmes d'Information et de Décision (INFORSID), pp. 93-108.

[15] Lejeune, C. (2014), Manuel d'analyse qualitative, De Boeck.

[16] Miner, G. & al. (2012). Practical text mining and statistical analysis for non-structured text data applications, Academic Press.

[17] Pénin, J. (2011). Open source innovation: Towards a generalization of the open source model beyond software. Revue d'économie industrielle, (136), 65-88.

[18] Simon, L. (2009). Underground, upperground et middle-ground: les collectifs créatifs et la capacité créative de la ville. Management international, 13, 37-51.

[19] Somers, S. (2009). Measuring resilience potential: An adaptive strategy for organizational crisis planning. Journal of contingencies and crisis management, 17(1), 12-23.

[20] Suire, R. (2016). La performance des lieux de cocréation de connaissances. Réseaux, (2), 81-109.

[21] Troxler, P. (2014). Fab Labs forked: A grassroots insurgency inside the next industrial revolution. Journal of Peer Production, 5.

[22] Utterback, J.M. (1994). Mastering the dynamics of innovation, Harvard University Business School Press, Boston Massachusets.

[23] Viseur, R. (2014). Automating the Shaping of Metadata Extracted from a Company Website with Open Source Tools, International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing.

[24] Viseur, R. & Charleux, A. (2021a). Open source communities and forks: a rereading in the light of Albert Hirschman's writings, Open Source Systems 2021.

[25] Viseur, R. & Charleux, A. (2021b). Contributions et coordination des makers face à la crise du COVID-19, Terminal, n°130.

From Open Science to Open Source (and beyond)

A Historical Perspective on Open Practices without and with IT

Bastian Wolff
The University of Cologne
wolff@wim.uni-koeln.de

Daniel Schlagwein
The University of Sydney
schlagwein@sydney.edu.au

ABSTRACT

Openness as organizational philosophy and theoretical concept has continuously gained importance over the past decades. While the adoption of open practices such as open-source development or crowdsourcing is primarily academically observed in the 20th and 21st century, organizational practices adopting or facilitating openness have already been applied before there was an understanding what openness actually depicts. For centuries, public and private stakeholders utilized a broad variety of open practices such as open science, industrial exhibitions, solution sourcing or industrial democracy in order to achieve certain anticipated effects – fully in the absence of IT. Due to the missing historical understanding, this paper provides a first holistic historical perspective on the emergence of open practices, considering the context of the political, technological and societal developments. Utilizing a structured literature review, the paper puts a special focus on the historical narrative and the connection between openness without and with IT.

The paper concludes that open practices are not a recent phenomenon, but were already applied successfully by affected stakeholders in previous centuries, whereas applied open practices partly build upon each other and show resembling patterns. Historically, two central shifts are identified: (1) a shift from government-driven towards organization- and community-driven open practices, and (2) a shift from mainly transparency-oriented open practices towards a stronger utilization of inclusion.

CCS CONCEPTS

• **Professional topics** → History of computing; • **Document types** → General conference proceedings.

KEYWORDS

Open Practices, Organizational Openness, Management History, Historical Narrative, Literature Review

ACM Reference Format:
Bastian Wolff and Daniel Schlagwein. 2021. From Open Science to Open Source (and beyond): A Historical Perspective on Open Practices without and with IT. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3479986.3479990

1 INTRODUCTION – OPEN PRACTICES AND THEIR HISTORICAL CONTEXT

Over the past decades, organizational openness has emerged as academic focal point, questioning and refuting the perception of organizations as fully closed entities without interfaces to their internal or external environment. Openness is associated with certain degrees of transparency and inclusion [85] with regard to organizational elements such as resources, participation processes or democratizing effects [69]. At this, transparency refers particularly to the visual or verbal visibility of these elements, while inclusion is associated with participation respectively involvement. Transparency and inclusion can act intra-organizationally or extra-organizationally [37, 85] and imply information exchange between involved stakeholders and entities, enabling organizational learning [55]. Different theoretical frameworks conceptualize openness, the most prominent being open innovation [17], open strategy [18, 85], open-source [66], open data [44], open education [77] or open government [47].

Nowadays, information technology (IT) plays a major role for organizations applying open practices. The majority of the observed contemporary open organizational practices such as open-source development [66], internal jamming [8], idea platforms [45] or crowdsourcing [1] are either purely IT-based or at least supported by IT. This can be also observed at recent cases such as Wikimedia [25], Daimler [57] or IBM [8], as well as in multiple sectors [65, 76]. Due to the fact that the academic focus on openness emerged particularly in the end of the 20th and begin of the 21st century (where IT has always been a pervasive factor) and that the majority of investigated cases also refers to the same era, the impression occurs that openness is a relatively new phenomenon and that, in turn, organizations were acting as closed entities before.

However, openness can be observed both with and without IT [69]. Including external elements to the own organizational setting respectively applying practices of democratization were already practices applied in times when IT was not the driver for organizational openness yet. Already in medieval and industrial ages, adopted organizational practices such as industrial exhibitions [15], inventing prizes [51], licensing external innovations [59], trade associations [6] or industrial democracy [81] did de-facto break the paradigm of organizations as closed entities which can be perceived as separated from their environment. In academia, open science emerged in the end of the 16th century as relevant philosophy, changing the way how researchers and scientists disclose their findings to the public and collaborate with their peers [21, 23]. All

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479990

these practices exhibit certain degrees of transparency and inclusion, despite of the fact that these practices were not necessarily defined as 'open' by practitioners or research.

It strikes that there are no dedicated holistic studies which deal with the historic utilization of open practices and their political, technological or societal context. Existing historic narratives are limited to a brief overview regarding historical manifestations of openness by Schlagwein et al. [69], revealing general historic milestones and their context. Besides, only industry- or subfield-specific narratives are performed, including openness in the computing and mobile phone industry [31] or in education [63]. In order to adjust the wrong perception of openness to be a phenomenon of the 20th and 21st century as well as to close this research gap, this paper intends to take a look back – providing a grand historic narrative on the emergence of open practices and their context.

Closing this research gap, the historic adoptions of openness are enlightened by the knowledge of organizational openness which exists today, enabling to identify trends regarding how open practices developed over time. Moreover, based on the findings regarding the historic development of open practices, an outlook can be derived which considers the context of contemporary open practices. Accordingly, the paper contributes to a better understanding of openness in the course of time as well as to a better understanding of open practices both without and with IT.

2 RESEARCH APPROACH

The paper utilizes a structured literature review, being considered particularly suitable in order to present the historical context and perspective of a respective phenomenon [35]. At this, the study considers the best-practice criteria and underlying principles for transparent historical narratives with regard to credibility, confirmability, dependability and transferability [32]. Due to the nature of the study, it applies an explorative hermeneutical (iterative, circulating) review and literature research approach without predetermined structure or topic setting [10]. With regard to literature search and study selection, the paper applies the recommendations by Kitchenham and Charters [52], particularly when it comes to inclusion or exclusion criteria.

We started the review by identifying and evaluating acknowledged fundamental publications such as Schlagwein et al. [69], David [23] or Ceruzzi [16] as well as the publication record of acknowledged business history journals (e.g. 'Business History Review') in order to establish a rough timeline regarding applied open practices. The further evaluation process was particularly focused on establishing a comprehension regarding historical proceedings, whether and how transparency and inclusion were utilized and what the relevant political, technological or societal background factors were. The iterative and circulating approach enabled the identification and evaluation of further publications, sharpening and verifying the picture of the individual open practices. In the case of doubts regarding the relevance or reliability of a publication, the content of the abstract, number of relevant citations as well as the reputation of the respective journal or conference were taken into account. Since several historic open practices are not labelled or known as 'open', relevant open practices were chosen based on the question if and how they made use of transparency and inclusion.

Due to the fact that many recent publications on openness are published as conference proceeding, the narrative considers both published journal papers and conference proceedings. Also, acknowledged academic books or further sources (e.g. working papers) are considered if they refer to the original source or provide further relevance. The review focuses particularly (but not exclusively) on academic publications due to the objective to provide an accurate grand narrative of open practices. At this, both historic publications as well as recent studies are taken into account. As basis for the literature review, the following data bases were used: ABI/INFORM, AIS eLibrary, EBSCOhost, JSTOR, ScienceDirect, Web of Science and Google Scholar. Relevant journals were partly accessed directly.

3 OPEN PRACTICES WITHOUT IT – BREAKING THE PARADIGM OF ACADEMIC AND INDUSTRIAL SECRECY

3.1 Open Science: Creating academic transparency

As one of the first structurally observed manifestations of openness, the phenomenon of open science emerged in the late 16th and early 17th century, starting a transition from concealing insights about nature's secrets towards disclosing new scientific knowledge towards peers and public. Overall, open science refers to "transparent and accessible knowledge that is shared and developed through collaborative networks" [79, p.434] and is embodied in many academic practices which we know today such as publishing academic insights to a wider audience (e.g. via open access), conducting peer-reviews, creating transparency regarding applied research methods or utilizing open knowledge repositories and data bases. However, for centuries scientists typically did not apply these practices of academic transparency which we know today. From the time when academia was founded in ancient Greece by (among others) Aristotle and Plato till the late Middle Ages and early Renaissance, the principles of secrecy in alchemy and science were upheld. Acknowledged scientists or inventors like Leonardo da Vinci (1452-1519) or Galileo Galilei (1564-1642) showed a very reluctant approach of showing their writings to others or even encoded their research results with anagrams, withholding their academic insights from the public [23].

The motives for this closed model of science, which sustained for so many centuries, are diverse: Despite of few exceptions, religious, political and societal norms contributed to withholding knowledge from the 'unworthy multitude' both in ancient Greece or medieval times [23]. Also, scientists often wanted to protect and control their knowledge in order to find a way to commercialize their ideas and inventions. The scientists' dependence from different kinds of patrons (providing political and material support) brought often the necessity of exclusivity: Only if findings were unique and prime, anticipated benefits of the scientist-patron relationship could be achieved [21]. The fact, that the whole academic philosophy and sheer idea of science came from a closed model, did not make it easier for scientists to apply more transparency – there was simply little understanding how open science could look like and how it could be advantageous. Consequentially, secrecy and informational asymmetries were normal conditions in the world of science.

However, while aristocratic patronage was initially a reason for scientists to sustain secrecy, it can be also perceived as one of the initial promoters of open science. For a long time, secular or religious patrons were historically closely related to European science. At this, nobles typically had two central motives to act as scientific patron: ornamental motives (focusing on self-aggrandizement, status or reputational advantages against other nobles) or utilitarian motives (focusing on the contribution of the research results for economy, military or society) [21]. For ornamental motives, it was elementary for inventions and academic results to be published, in order to be publicly recognized for the achievements and discoveries. For both scientists and nobles, demonstrating knowledge and inventions to the outside world brought a notable raise in societal and scientific legitimacy. However, many utilitarian inventions (e.g. inventions providing superior geographic knowledge regarding important trade routes) demanded a certain level of secrecy [21, 23].

Hence, the question arises, how the patrons' motives contributed towards a more open form of science from the post-renaissance era? The answer is closely related to the implications of the previously mentioned information asymmetries. The more intense use of mathematical methods of post-renaissance scientists and their drive to reveal the 'unfamiliar' created a certain dilemma for potential patrons: Evaluating academic methods and insights was intellectually far more demanding than before, given the lack of specialized expertise among the patrons. In order to protect themselves from embarrassment through sponsoring 'charlatanry', they demanded peer-reviews which could only be conducted by other mathematical-oriented scientists [21-23]. Along these developments, so-called 'cooperative rivalries' between scientists emerged as a functional response to the existing information asymmetries, building up on the insight that the disclosure of knowledge, demonstrating inventions and open inquiries contribute to the generation of new knowledge [22].

The occurring re-organization of European science could be particularly observed in the form of open practices such as "participation in informal networks of correspondence, [. . .] public challenges and contests, open demonstrations and exhibitions, and the certification of individuals by cooptation and election to 'learned societies' " [21, p.578]. These new practices profited strongly from emerging technologies and regulations such as copyright privileges, postal dispatch systems or Guttenberg's printing press, forming background conditions for diffusion and protection of knowledge alike [23]. The London 'Invisible College' from 1646 depicted one of the first scientific groups representing these aforementioned 'informal networks': Through informal meetings and letter exchange dealing with their scientific insights they thrived on the fundamental idea that the open knowledge exchange contributes to the creation of new knowledge [20]. As one of the consequences, the 'Philosophical Transaction of the Royal Society' introduced the first peer-reviewed journal in 1665, building a milestone regarding academic transparency [9].

3.2 Industrial Exhibitions: Publicly showcasing inventions and industrial goods

Comparable to the medieval alchemists and scientists, also the European crafts guild considered their specific knowledge and technological assets worth protecting. While journeymen contributed to the diffusion of industrial craft knowledge on a local level, and first notable medieval trade fairs like the 'Messe Frankfurt' emerged at the hubs of the trade routes, broader transparency was often prevented by economic interests and regulatory factors [23, 68]. However, the rise of open science brought along practical implications for the industrial development of Europe and the way how it showcases technology and new inventions to the public. What we understand nowadays as 'economy' or 'industry' was in pre-industrial times closely connected to the interplay between academia, inventors and patrons – many of the economic-relevant inventions were made by scientists and applied in the non-academic world, particularly after the rise of mathematical methods in science. Accordingly, open science does not only refer to openness in academia which can be separated from economic technological research. As a consequence of the economic- and reputation-driven patronage which facilitated open science and its overarching socioeconomic impact, the emerging open practices regarding public contests, demonstrations and exhibitions provided a basis for the further development of open practices, also in the area of industrial discoveries [21].

Evolving from the aforementioned medieval trade fairs and the new academic exhibition practices, industrial exhibitions provided an opportunity for firms and inventors to exhibit new inventions, technology and industrial goods to a larger audience of interested stakeholders, potential buyers or investors. Taking place since the 17th century (the industrial exhibition in Paris 1683 being one of the first documented organized industrial fairs), the sporadic exhibitions in these early days gained usually relatively little public attention [15, 27]. However, they provided a basis for bigger industrial transparency and established themselves as mass phenomenon in the 19th century (one of the highlights being the international 'Great Exhibition' in London 1851), profiting from the technological advances of the industrial revolution [2].

Resembling the emergence of open science and its motives, also industrial exhibitions were strongly driven by nobles and governmental stakeholders. At this, the exhibitions had multiple objectives: Governments and nobles aimed on facilitating the development of new relevant technologies, increasing the national prestige or stimulating exports, why the expositions were often organized or financed by according stakeholders. As an example, the French government introduced official industrial exhibitions at the very beginning of the 19th century in order to react on the rise of British products which were produced more efficiently as a consequence of the industrial revolution [34]. Exhibitioners on the other hand participated for the monetary value of the exhibition prizes, local pride or the possibility to promote their inventions to a wider audience [15, 34]. As consequence, trade fairs and industrial exhibitions contributed to a switch from industrial secrecy towards more transparency with regard to industrial goods and inventions, leading to more openness both on the economic and organizational level. This is also observed by Landes [53] and Dunham [26], who emphasize the role of exhibitions in diffusing knowledge and technology.

3.3 Inventing Prizes: Between solution sourcing and fostering economic development

Often applied as part of the industrial exhibitions, monetary exhibition prizes were provided as incentive and reward for the best exhibition pieces [34]. However, in the 18th and 19th century this practice was further adapted by governmental entities and organizations in order to source for external solutions and innovations. Inventing prizes and innovation awards induced inventors to deal with a specific problem or subject and reveal their inventions and ideas, facilitating openness. Beside of the offered incentives, the possibility to further commercialize the invention through patents or overlapping awards made it often further attractive for inventors to participate [51].

At this, two main approaches could be observed: The first one being the utilization of inventing prizes in order to source for a specific solution to a particular problem. As popular example, the British parliament publicly announced a notable reward for finding an accurate way of measuring the longitude at sea in 1714. While prizes were particularly promised by national states and other government entities, also private entities such as the billiard table producers Phelan and Collender, who were looking for a cheaper material to produce billiard balls in 1863, were using promised incentives to approach specific organizational challenges. Similar cases were observed all over Europe. Hence, this approach provided certain elements of inclusion through involving external parties into the own solution finding and R&D process [51].

The second form of utilizing inventing prizes was less focused on finding a solution to a particular existing problem, but rather to generally foster economic development by promoting new innovations and overcoming information asymmetries through the diffusion of technological knowledge. Increased transparency of the inventions was an anticipated side-effect, creating the wanted spread of information and innovation for the sake of economic development. Private or governmental prize-granting institutions such as the British RSA (Royal Society of Arts; founded in 1754) or the French SEIN (Society to Encourage National Industry; founded in 1801) are representative entities following this objective [51]. In this regard, the motives for inventing prizes are closely connected to the motives of industrial exhibitions.

3.4 Solution Sourcing: Utilizing external innovations via patents and licenses

The first described approach regarding utilizing inventing prizes (focusing on sourcing for special solutions) was not necessarily a practice applied only in the context of inventing prizes. Craftsmen, merchants and organizations were constantly buying or copying new external inventions in order to apply them on their own. At this, the adoption of new external innovations depicted a logical necessity: If a competitor had a new and better way how to produce or transport a certain good, adaption respectively the integration of new technology was often the only way how to stay in competition. For instance, when the modern magnetic compass emerged in Europe around the 11th/12th century (it is questionable if it came via the sea routes from China respectively Arabia, or was invented independently), the novelty was quickly utilized by merchants and explorers, despite of the fact that these stakeholders had not necessarily invented it [72]. In the absence of explicit regulations, new inventions could just be utilized as soon the knowledge about the invention was spread via word-of-mouth or if somebody bought it directly from an inventor. However, as long there was no possibility for inventors to protect their inventions and intellectual knowledge, revealing novel inventions included the danger to potentially lose the possibility of future monetization.

With the emergence of modern patent and intellectual property law, also the practice of utilizing external innovations evolved. Inventors or organizations could protect their inventions in order to use them themselves, sell licenses to others to further commercialize the patents, or sell the patents. Patent and copyright laws go back to the 15th century, where the Venetian Republic in the effort to attract skilled artisans and inventors was offering exclusive rights for their inventions. Being already an important background factor for open science, the fundamental idea of copyrights and patents was adapted around Europe in order to facilitate economic development. In the US, the first article of the constitution (1789) assured inventors exclusive rights to their inventions [59].

As a consequence of the industrial revolution, the ongoing mechanical progress and the possibilities for inventors to protect their intellectual knowledge, more and more inventions were patented. At this, independent inventors would not necessarily open own shops (also due to the lack of resources) but sell their patents or provide licenses to one or more manufacturers [60]. Organizations were faced with the decision to either develop and manufacture industrial technology and goods themselves or to purchase the external invention from private persons or other organization (e.g. via buying or licensing a patent or contracting the production of industrial good in the form of an outsourcing or procurement agreement). For many firms, like the Draper Company which invented the Northrop loom (becoming the industry's standard in 1895), defending their respective patents and market them in a smart way was key for their economic success – other companies were required to enter licensing agreements to stay competitive [56]. Overall, these forms of external solution sourcing showed certain characteristics of inclusion through involving external stakeholders and their inventions to the own organizational setting.

3.5 Trade Associations & Cartels: Formalized cross-organizational aligning

While open practices like open science or industrial exhibitions focused strongly on increasing transparency and were driven by governmental stakeholders, occurring cross-organizational alignments were often driven by self-interest of organizations, merchants and craftsmen. The history of modern formalized cross-organizational alignments started already in medieval times: Local or alien guilds of merchants and craftsmen used their mutual market power in order to fix prices, lobby at local authorities, restrict supplies and control market entries, hence, enforcing monopolies and oligopolies. So from a pure business model or market entry perspective, the guilds did not promote openness, but rather suppress competition. However, the guilds applied cross-organizational openness through revealing e.g. prices or operating principles to other members (transparency) and surrendering decision power under mutual regulations for the common benefit (inclusion). At this, the guilds were generally quite restrictive – internal punishment for violating the guild's regulations assured the integrity of the respective guild [61]. As acknowledged example, the 'Hanse' was formed in the 12th century as commercial association of northern German merchants, in order to establish trade route protection and formulate common economic interests. While it existed up until the 17th century, it is commonly considered to be one of the first established trade associations [54].

The modern trade associations emerged in the mid of the 19th century, also as a consequence of the very tight policies of their predecessors and the further institutionalization of companies. While these trade associations had various motives, stricter anti-trust legislation caused a shift towards more individualism within the associations, allowing an enhanced degree of individual business practices and free competition [58]. But still then, these modern trade associations were often acting as a tool in order to create different forms of price cartels and suppress competition: At the end of the 19th century, US hardware wholesalers utilized trade associations in order to stabilize prices and increase the negotiation power against other complementary trade associations [4]. At this, including the other companies into stating prices and negotiating brought two main benefits: First, information asymmetries, which naturally existed with regard to other firms' prices, were massively reduced. Second, a single company profited from the resources of the network when negotiating with other stakeholders. With regard to both, certain degrees of transparency (e.g. in prices, market positioning, cost of operations) were a key element in order to benefit from the anticipated advantages. With regard to inclusion, companies transferred certain decision and negotiation power to the trade association which, in turn, was then able to establish the strategic positioning of the association.

However, associations were not only acting as a tool for organizations to establish cartels: Berk and Schneiberg [6] observe a development of American industry associations from being cartels to being developmental associations from 1900-1925: In order to achieve organizational learning effects, collaborative associations, committees and deliberative forums were providing a place for representatives to learn from and discuss with external stakeholders from other associations, organizations, governments or agencies. At this, the discussions and thought experiments were facilitating knowledge transfer and information sharing, very often regarding costs and productivity – creating a competitive advantage against non-participating organizations. In comparison to the usage of associations as tool in order to exclusively form cartels and basically suppress competition, the information transparency in these developmental associations is much more related to organizational learning: Using external knowledge in order to improve own production processes and strategic considerations.

3.6 Industrial Democracy: Establishing internal participation and democratization

Till the 19th century, open practices were fairly focused on extra-organizational open practices, being oriented on increasing transparency, utilizing external resources, knowledge or inventions, or establish cross-organizational alignment. Intra-organizational open practices in the form of democratization, common decision making or bottom-up involvement were not applied on the broad scale. While several contemporary open practices involve employees to decision making, such as consultative participation, employee ownership, representative or informal participation, work councils, board level representation or social media jams [29], this was not the case for a very long time. Although several cases of joined negotiations or even what we understand by trade unions can be identified along history [82], craftsmen and journeymen only very occasionally joined forces in order to stand united against their employer in order to demand better wages, working conditions or involvement into decision making.

The emergence of industrial democracy at the begin of the 19th century depicted a paradigm shift towards a stronger (internal) involvement of factory workers. Industrial democracy, a term created and shaped by Webb and Webb [81], refers to employees' involvement into decision making or collective bargaining, employee representation or further types of employee empowerment. In order to comprehend the emergence of industrial democracy, one must consider the industrial world's context of the early 19th century: The rapid expansions of the first industrial revolution – facilitated by breakthrough inventions like the steam engine or the power loom – brought huge implications for workers. Previously hand-crafted products could be now produced with mechanical help, changing their job profile towards repetitive work as well as raising the need for more unskilled or semiskilled workers. Child and women labor rose drastically, similar to the need for coal or iron miners [40, 41]. Consequentially, the employer-worker relation in industrial organizations in Europe or the USA was characterized by wage-labor, long working hours, strong hierarchies and little involvement of workers regarding internal participation or democratization. Work conditions, contracts, salaries as well as decision making was mainly in the hand of management respectively owners, resulting in strong power asymmetries [24, 40].

Under these circumstances, several societal and political impact factors are considered to have led to more industrial democracy. Weighing up existing historical investigations, Hyman [46] identifies a multi-step process to have consequentially led to the emergence of industrial democratization, consisting out of (1) the achievement of political democracy, (2) the accompanying impact on workers with regard to ideas on social democracy and representation, (3) the demand that this social level of democracy and voice also accounts at the work place and (4) the overarching impact of economic democracy, leading to organized workers' unions, shop committees or even self-governing workshops. In Europe, this 'politicization' of the factories was also facilitated by arising socialist/communist (e.g. Karl Marx or Friedrich Engels) respectively liberal (e.g. John Stuart Mill) thought leaders who questioned the power distribution in the factories [19, 46] as well as the described negative working conditions. Also, certain economic factors played a relevant role: In the USA, declining economic conditions led to a mutual dependency between employers and workers, which further supported the formation of common employer-worker committees. Examples of these employer-worker committees are observed in the US woodworking industry, where these committees elaborated and published in a cooperative effort price books and quality standards for new products during the 1820's [33].

Although employers and governments often reacted harshly and repelling to first structured formations of shop committees or trade unions (e.g. at the formation of the 'Grand National Consolidated Trades Union' in the UK in 1834), internal participation and democratization established themselves as practices in the industrial world [24]. With the further growing importance of trade unions and shop stewards in the 19th and 20th century, also employee representation became more important with regard to collective bargaining when it came to protect working practices, receive better wages and improve existing working conditions [19].

4 OPEN PRACTICES WITH IT – THE COMPUTERIZATION OF OPENNESS

Until mid of the 20th century, openness by transparency or inclusion happened purely in the absence of information technology (IT). This does not mean that technology did not facilitate openness before: As indicated, breakthrough inventions such as the printing press, postal services, telegraphy or new transportation technology were used before in order to promote open practices by enabling diffusion of knowledge or enabling inclusion of other stakeholders [23, 87]. However, physical limitations for involved stakeholders, resources or processes always provided certain restrictions regarding potential anticipated outcomes. The exchange of information and the interaction of affected stakeholders, which are both embedded in the very nature of openness, was always limited to the technological possibilities of the respective time. With the emergence of commercial computers in the 1950's, also the history of applied open practices entered a new era by overcoming step-by-step these physical limitations. The newly created context of physical hardware devices (e.g. mainframe computers or later PCs and mobile devices) and particularly software (e.g. applications, data bases or operating systems) provided possibilities which enabled the opening of organizational elements via new ways of coordination or facilitation.

4.1 Open-Source Development: Opening software's source code

World War II, which facilitated advances in technological areas like code breaking activities, electronic calculations or material research as well as produced skilled engineers for the civil market, created a broad basis for the emergence of the modern age of computerization, particularly in the United States. Consequently, computers like the 'UNIVAC', which was released in 1951 by the Eckert-Mauchly Computer Corporation, or IBM's more successful '701' from 1952, started the effort to manufacture and distribute commercial computers, offering the promise of huge speed advantages compared to non-electronical calculators [16]. Along with the new possibilities, commercial use cases were identified in all possible fields such as in data analytics [48], biological taxonomy [74] or mechanical engineering [64]. While users appreciated the new possibilities provided by computers, there was no clear differentiation between hardware and software to that time. Hardware and software were typically provided by the same supplier. The code for the software itself was usually accessible and changeable for the users, who simultaneously were also acting as programmers. Accordingly, open-source code was an early reality. This also brought along first cross-organizational collaborations like PACT (Project for the Advancement of Coding Techniques), where software engineers of multiple companies used the open code and mutually programmed a shared set of tools in order to create common value [83].

With time and new emerging use cases, more and more programming languages such as FORTRAN or COBOL emerged, providing possibilities for companies to code on a higher level. However, with the software environment getting more complex and diversified and due to the nature of many emerging compilers (which translated the software's source-code into binary computer-readable machine code), possibilities for computer and software suppliers emerged to only release the binary code of a software, making it difficult for other programmers to read or use it. As a result, the period from the 1960's till 1980's brought an increasing number of stand-alone software products with closed source code, Microsoft being a well-known representative of this development [16, 83].

However, the principles of open-source software were upheld by the IT community itself: In 1969, Ken Thompson and Dennis Ritchie from the Bell Telephone Laboratories (the former research department of AT&T) started to develop the UNIX operating system. Up until the 1980's, UNIX was particularly used at US universities as open-source solution. Although it was commercialized by AT&T in the 1980's, it had remarkable impact on the programmers' community and provided a basis for future open-source operating systems [16]. Also, in response on the trend towards closed software, Richard Stallman (an acknowledged programmer from the MIT) started the 'free software movement' in 1985, focusing on establishing a legal and practical framework for free access to software and its source code. The movement built up on the idea that software authors could use copyright and licensing law in order to preserve the status of their software to be 'free' [39]. This community-based idea was shared among relevant parts of the IT community: open-source-based operating systems such as GNU or BSD evolved in the 1970's and 1980's as part of a practitioner movement of software programmers and engaged communities [13, 83].

Bigger achievements were particularly accomplished in the 1990's with the development of the Linux open-source operating system by Linus Torvalds [13] or of the open-source data base MySQL by Michael Widenius and David Axmark [86], still being today among the most popular solutions of their kind. The aforementioned 'free software movement' also provided the intellectual and legal basis for Bruce Perens and Eric Raymond who did start the so-called 'open-source software movement'. Building up on similar legal licensing principles, the 'open-source software movement' emphasizes the actual commercial and practical benefits of open-source software [62]. In his acknowledged conference submission 'The Cathedral and the Bazaar', Raymond [66] provides a baseline idea how to perceive open-source and which benefits publicly available source code provides. Since then, the term 'open-source' is used in the academic and professional context. While the emergence of open-source solutions was particularly driven by engaged communities, a switch towards a more commercialized approach could be observed in the 2000's, being characterized by a stronger emphasis on product delivery and support [28]. Remarkably, the adoption of open-source software practices had also relevant impact on inner-organizational openness initiatives: Labelled by Tim O'Reilly in 2000, 'Inner Source' brought open-source principles such as open communication, open development artifacts (e.g. source code) or open collaboration into organizations, facilitating inner-organizational transparency and inclusion via open-source practices and culture [14].

4.2 Crowdsourcing, Jamming, Idea Platforms and Co.: The rise of internet- and intranet-enabled open practices

The rise of the internet from the mid of the 1990's brought along a variety of new possibilities how people collaborate and exchange knowledge. Among others, it provided a further boost for the open-source movement with its ability to simplify the sharing and accessing of source code as well as enabling collaborations with low transaction costs for involved stakeholders [5, 78]. However, the emergence of internet respectively intranet technologies and the simultaneous rise in wide-ranging private and professional IT device usage (also due to cheaper prices and better user interfaces) had much broader implications: It both enabled a much broader involvement of contributors as well as recipients. Suddenly, not only a limited group of internal employees could be involved into development or content creation, but everybody who had a device with an intranet or internet access. Also, the transaction costs for potential contributors were reduced drastically: Accessing, sharing and collaborating via the own computer was often cheaper and less coordination- or time-intense than physical person-to-person alignments. Accordingly, the possibility to access quickly all kinds of available knowledge made the internet a melting pot for entrepreneurs, knowledge seekers and ordinary people alike.

The consequence was a boost for various open practices which built on the principle of mass participation. The case of Wikipedia (which was founded in 2001 as free internet encyclopedia) shows, how the internet facilitated openness and freedom for involved (external) stakeholders when it came to involvement, access and control of the content [67]. IT had become a tool and facilitator to support openness, enabling involvement of people who were not necessarily IT-affine before. This can be also observed at the practice of online crowdfunding campaigns (raising money in order to finance a certain purpose or project), where the internet in combination with innovative IT solutions enabled open inquiries and open contributions by private or organizational stakeholders [38].

The advantages of the networked world and the reduction of the transaction costs for involvement also enabled organizations to approach the more IT-affine part of the internet community by considering their contributions in crowdsourcing initiatives: As solution-oriented alternative to internal sourcing or utilizing a supplier, crowdsourcing (referring to the inclusion of mostly external crowds into solution, idea, content or product generation) acted as further sourcing option regarding bringing external knowledge, skills, solutions or unbiased opinions into the organization in order to solve certain problems in a collaborative effort [1]. When Howe [43] labelled crowdsourcing as a term in 2006, the practice was already commonly applied in multiple IT-related cases. At this, people engaging in crowdsourcing campaigns are largely professionals and experts, while the received recompense from the crowdsourcing arrangements is usually small compared to the invested work and expert knowledge, making it particularly attractive for companies to make use of the crowd [11]. A specific sub-type of these popular expert-directed crowdsourcing campaigns are the so-called bug bounty programs, representing a modern form of the previously introduced inventing prizes. Gaining broader relevance in the 2010's, organizations and software developers challenge external experts from the hacker community to identify security-related software bugs and vulnerabilities, often under the promise of certain incentives [75].

Intranet and platform technologies also contributed to various emerging intra-organizational open practices: In a world-wide (150.000 employees from 104 countries) internal project, IBM conducted in 2006 an internal 'innovation jam' which was performed in two 72-hour sessions. Interlinked bulletin boards and intranet pages enabled internal cooperation and coordination in order to facilitate brainstorming and idea generation [8]. Also internal [45] or external [42] idea platforms emerged, encouraging the involvement of staff or customers into product development, knowledge exchange or common decision making.

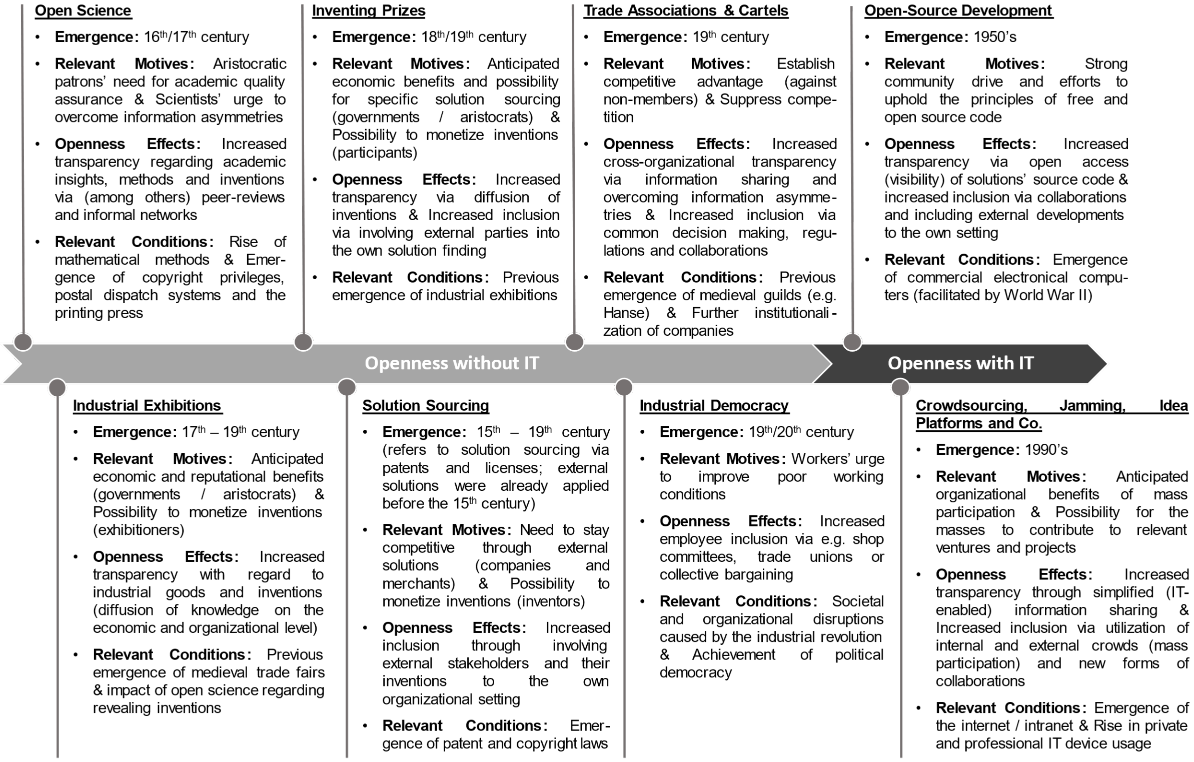
5 TRENDS, OUTLOOK AND RESEARCH LIMITATIONS

The continuous emergence of open practices along the centuries brought groundbreaking and defining changes towards how we understand organizations today – be it practices like open science, exhibitions or trade associations which still can be observed today, the lasting implications of industrial democracy or the pervasive open practices with IT which still shape the economy more than ever. Freeman [30, p.40] observes that industries continuously co-evolve in a complex interrelated process "between science, technology, economy, politics and culture" – and so does openness. Figure 1 provides a simplified timeline illustrating the historical emergence of open practices.

Analyzing the historic timeline, the narrative reveals certain trends and shifts which occurred along the centuries, providing a basis how open practices can be perceived in the context of the times.

5.1 The Shift from Government-driven towards Organization- and Community-driven Open Practices

As indicated by the historic narrative, different stakeholders were adopting or facilitating open practices in order to utilize them for their own motives. When it comes to the driving forces behind their emergence, research claims that early open practices such as open science, industrial exhibitions or inventing prizes were often driven or at least promoted by governmental stakeholders and nobles with



Open Science

- **Emergence:** 16th/17th century
- **Relevant Motives:** Aristocratic patrons' need for academic quality assurance & Scientists' urge to overcome information asymmetries
- **Openness Effects:** Increased transparency regarding academic insights, methods and inventions via (among others) peer-reviews and informal networks
- **Relevant Conditions:** Rise of mathematical methods & Emergence of copyright privileges, postal dispatch systems and the printing press

Inventing Prizes

- **Emergence:** 18th/19th century
- **Relevant Motives:** Anticipated economic benefits and possibility for specific solution sourcing (governments / aristocrats) & Possibility to monetize inventions (participants)
- **Openness Effects:** Increased transparency via diffusion of inventions & Increased inclusion via involving external parties into the own solution finding
- **Relevant Conditions:** Previous emergence of industrial exhibitions

Trade Associations & Cartels

- **Emergence:** 19th century
- **Relevant Motives:** Establish competitive advantage (against non-members) & Suppress competition
- **Openness Effects:** Increased cross-organizational transparency via information sharing and overcoming information asymmetries & Increased inclusion via common decision making, regulations and collaborations
- **Relevant Conditions:** Previous emergence of medieval guilds (e.g. Hanse) & Further institutionalization of companies

Open-Source Development

- **Emergence:** 1950's
- **Relevant Motives:** Strong community drive and efforts to uphold the principles of free and open source code
- **Openness Effects:** Increased transparency via open access (visibility) of solutions' source code & increased inclusion via collaborations and including external developments to the own setting
- **Relevant Conditions:** Emergence of commercial electronical computers (facilitated by World War II)

Industrial Exhibitions

- **Emergence:** 17th – 19th century
- **Relevant Motives:** Anticipated economic and reputational benefits (governments / aristocrats) & Possibility to monetize inventions (exhibitioners)
- **Openness Effects:** Increased transparency with regard to industrial goods and inventions (diffusion of knowledge on the economic and organizational level)
- **Relevant Conditions:** Previous emergence of medieval trade fairs & impact of open science regarding revealing inventions

Solution Sourcing

- **Emergence:** 15th – 19th century (refers to solution sourcing via patents and licenses; external solutions were already applied before the 15th century)
- **Relevant Motives:** Need to stay competitive through external solutions (companies and merchants) & Possibility to monetize inventions (inventors)
- **Openness Effects:** Increased inclusion through involving external stakeholders and their inventions to the own organizational setting
- **Relevant Conditions:** Emergence of patent and copyright laws

Industrial Democracy

- **Emergence:** 19th/20th century
- **Relevant Motives:** Workers' urge to improve poor working conditions
- **Openness Effects:** Increased employee inclusion via e.g. shop committees, trade unions or collective bargaining
- **Relevant Conditions:** Societal and organizational disruptions caused by the industrial revolution & Achievement of political democracy

Crowdsourcing, Jamming, Idea Platforms and Co.

- **Emergence:** 1990's
- **Relevant Motives:** Anticipated organizational benefits of mass participation & Possibility for the masses to contribute to relevant ventures and projects
- **Openness Effects:** Increased transparency through simplified (IT-enabled) information sharing & Increased inclusion via utilization of internal and external crowds (mass participation) and new forms of collaborations
- **Relevant Conditions:** Emergence of the internet / intranet & Rise in private and professional IT device usage

Figure 1: The Emergence of Open Practices – Simplified Historical Timeline. (Own Figure).

economic and reputational motives [15, 23, 34, 51], indicating the strong role of patronage, aristocracy and governmental influence on scientific and economic developments. The necessary networks and monetary resources which were used in order to facilitate these open practices were often occupied by nobles or governments, making it easier for them to follow their particular motivations. Despite of the fact that on a smaller scale also very early organization-driven open practices are observed such as early forms of solution sourcing or the formation of the local guilds, the structural capabilities and available resources of inventors or companies did often not provide the possibility for broad-scale openness. The described information asymmetries between involved stakeholders did certainly not make it easier for them to exchange information, considering the danger to lose competitive advantages. Hence, it is indicated that early forms of structured open practices were fairly often promoted by nobles or governments.

On the contrary, the narrative implies that later open practices without IT such as trade associations or industrial democracy exhibit a higher degree of self-motivation of the organizational stakeholders. While governmental stakeholders still influenced regulatory and economic conditions, the further formal institutionalization of companies as corporations contributed to competitive situations and organizational structures which facilitated the emergence of organization-driven open practices. In this context, it is argued that what we understand as modern corporations emerged from the 17th till the 19th century, which included huge implications regarding structural changes and economic power [49]. The original motives of nobles and governments to promote open practices (economic development and reputational effects) were partly substituted by motives of organizations to apply openness for their own sake, respectively by employees who strived for democratization. Particularly the community-driven emergence of open practices with IT such as open-source development finalizes this shift: In many cases, IT has become an enabler for self-driven engaged communities facilitating open practices. However, while the community-factor plays a huge role in almost all contemporary IT-enabled open practices, a clear differentiation between community-driven and organization-driven open practices is hardly possible or even sensible – organizations, communities and individuals are closely linked and so is the drive behind their actions.

Accordingly, this shift must be interpreted with care since historic mechanisms are highly complex. Ultimately, the driving forces and motivations behind emerging open practices cannot be comprehended independently from the historic dynamics of the contextual technological and societal developments. These dynamics have reciprocal implications on how people and organizations communicate, how organizations are structured and how power is distributed among involved private, organizational and governmental stakeholders. Hence, stakeholder motivations and actions cannot be evaluated in an absolute manner.

5.2 The Shift from Transparency towards Inclusion

Similar to the shift of the driving stakeholders, the narrative indicates a historical shift from open practices utilizing transparency towards utilizing inclusion. Early open practices such as open science or industrial exhibitions were particularly focused on information sharing via presenting new insights, inventions or products [23, 34]. Similar to the observation with regard to the shift towards more organization- and community-driven open practices, inclusion in the form of participation at cross-organizational alignment (trade associations), decision processes (industrial democracy) or product development (solution sourcing) became a bigger factor in the later era of openness without IT and particularly at openness with IT. The fact that the rise in inclusion went along with a rise in organization- and community-driven openness indicates that inclusion is an element which is closely connected to a self-driven motivation – collaboration, involvement of stakeholders and democratization imply a certain element of active contribution which typically requires motivation to act accordingly.

At this, IT (particularly the internet/intranet) as enabler of inclusion shines out due to its ability to enable mass participation for practices like crowdsourcing, innovation jams or idea platforms. It strikes that many open-source projects emerged as part of programmers' movements. All IT-related open practices have a very strong community character, emphasizing the social aspect of the collaboration. This reveals a certain insight: Openness with IT lives from engaged people and communities – openness becomes a social practice [73]. In this context, the resembling patterns of certain open practices without IT and open practices with IT exemplify the inclusion-enabling role of IT: Solution sourcing (without IT) and crowdsourcing (with IT) build up on the same fundamental principle to include external resources like inventions or knowledge to the own organizational setup. Industrial democracy with its fundamental idea of codetermination and workers empowerment (without IT) resembles on a smaller scale and in a different context what we can observe today with bottom-up approaches like innovation jams or internal idea platforms (with IT). Inventing prizes (without IT) and bug bounty programs (with IT) both utilize incentives in order to reveal and solve certain problems with the help of external stakeholders. Accordingly, the utilization of IT in the form of digitized practices enables enhanced inclusion-potentials when it comes to factors like range, participation outcomes, location and time flexibility [36]. To that regard, IT enables a much simpler and easier access to relevant resources (such as code) and lowers transaction costs for participation. Whittington [84] refers here to the 'massification of strategy', raising the fact that mass-produced hardware, software tools and connectivity facilitate the strategic participation of people beyond the hierarchical elites. Hence, also the transaction costs for inclusion drop, which further facilitates inclusive practices.

5.3 Outlook – Future Indications for Open Practices

The historic narrative has shown how open practices continuously evolved up into the 21st century. With regard to the continuation of the history of open practices and related further research, Hautz et al. [37] illustrate that new forms of open practices are constantly emerging along with the development of new types of information technology – an insight which is also indicated by this paper. The interplay of this ongoing technological progress with the current societal and economic disruptions which e.g. go along with the COVID-19 pandemic will certainly be a major research field with regard to the historic development of open practices – Particularly since the current COVID-19 related developments indicate that uncertainty and the inter-connectivity of industries with their environments are challenges which potentially affect all sectors [3] and imply extensive social and economic consequences for individuals and organizations alike [12].

While disruptive events of this scale affect the diffusion of knowledge as well as how stakeholders interact with each other or with technology, emerging IT-based or IT-supported open practices could potentially reflect this change. This paper has shown how open practices emerge in the context of political, technological or societal developments along the centuries. Current ongoing economic meta-trends such as platform economy [50], intra- and extra-organizational social media usage [7] or work flexibility [80] could act as a further facilitator of open practices. Accordingly, future research could investigate how open practices emerge as a consequence or under the impact of these ongoing societal disruptions and meta-trends.

Also, the emergence of new open practices without IT would be a particular interesting field for further research. While this paper indicates that the emergence of IT depicted a paradigm change for open practices, this does not exclude the possibility for new open practices which do not use IT as facilitator. New forms of inter-organizational strategy workshops [71] or of local open innovation labs [70] do not necessarily build up on IT which enables (virtual) mass participation. On the contrary, physical attendance might be an important anticipated factor in a world which has become more and more connected by IT. Investigating these (physical) open practices in the context of the societal disruptions resulting from the COVID-19 pandemic (which currently promotes a stronger virtualization) and the mentioned meta-trends, it might be interesting how these practices adapt, if they disappear, or if organizations actively promote them in order to achieve certain anticipated outcomes.

5.4 Research Limitations

It should be stated that this study exhibits certain research limitations: The narrative does not claim to consider all impact factors or to reveal all possible interrelated mechanisms which played a role in the emergence of the open practices. Historical mechanisms between political, societal, economic or technological factors are highly complex, creating always questions like "What caused what?", "How strong was the impact of factor a) on phenomenon b)?" or "How did factor a) impact phenomenon b) under the circumstance of c)?" – Revealing causal relations is an intricate field. Hence, the paper targets clearly on the logical narrative and the key mechanisms, without intending to provide an in-depth overview. Moreover, the study focuses particularly on Western regions (Europe and USA) when it comes to the historic emergence, neglecting e.g. developments in Asian countries.

6 CONCLUSION

Overall, the historic narrative reveals that open practices are certainly not a recent phenomenon of the 20th and 21st century, but have been applied by organizations and individuals since centuries: Open practices which facilitate transparency or inclusion are perpetually utilized by organizations and practitioners as consequence of the ongoing political, technological and societal developments. With the emergence of IT as pervasive socioeconomic factor of the 20th century and the later emergence of internet, intranet and platform technologies, also open practices entered a new era towards what we understand today by the term 'openness'.

The historic timeline reveals two central shifts, the first one being related to the driver behind the open practices: While early open practices such as open science or industrial exhibitions were strongly government-driven, organizations and individuals emancipated and formally organized themselves along the centuries, resulting in more organization-driven and (IT-related) community-driven open practices. The second identified shift refers to the change from mainly transparency-oriented open practices towards a stronger focus on inclusion, being rooted in an interplay of changing motivations and emerging technological possibilities. At this, particularly the emergence of IT and internet/intranet technologies acted as facilitator for inclusive practices due to the way how they enabled access to and diffusion of knowledge through the masses as well as their connective social nature – Open practices have taken a long road, emerging around a circle of continuous change and adaption. Considering the interplay of the ongoing development of new technological solutions and its societal and economic implications, there is no doubt that this story is going to be continued.

REFERENCES

[1] Allan Afuah and Christopher L Tucci. 2012. Crowdsourcing as a solution to distant search. Academy of Management Review, 37, 3 (2012), 355-375.
[2] Jeffrey A Auerbach. 1999. The Great Exhibition of 1851: a nation on display. Yale University Press, New Haven and London.
[3] Scott R Baker, Nicholas Bloom, Steven J Davis and Stephen J Terry. 2020. Covid-induced economic uncertainty. NBER Working Paper No. 26983, National Bureau of Economic Research, 2020.
[4] William H Becker. 1971. American wholesale hardware trade associations, 1870-1900. Business History Review (1971), 179-200.
[5] Yochai Benkler. 2002. Coase's Penguin, or, Linux and "The Nature of the Firm". Yale Law Journal (2002), 369-446.
[6] Gerald Berk and Marc Schneiberg. 2005. Varieties in capitalism, varieties of association: Collaborative learning in American industry, 1900 to 1925. Politics & Society, 33, 1 (2005), 46-87.
[7] Hardik Bhimani, Anne-Laure Mention and Pierre-Jean Barlatier. 2019. Social media and innovation: A systematic literature review and future research directions. Technological Forecasting and Social Change, 144 (2019), 251-269.
[8] Osvald M Bjelland and Robert Chapman Wood. 2008. An inside view of IBM's 'Innovation Jam'. MIT Sloan Management Review, 50, 1 (2008), 32-40.
[9] Bo-Christer Björk and David Solomon. 2013. The publishing delay in scholarly peer-reviewed journals. Journal of Informetrics, 7, 4 (2013), 914-923.
[10] Sebastian K Boell and Dubravka Cecez-Kecmanovic. 2014. A hermeneutic approach for conducting literature reviews and literature searches. Communications of the Association for Information Systems, 34, Article 12 (2014), 257-286.
[11] Daren C Brabham. 2012. The myth of amateur crowds: A critical discourse analysis of crowdsourcing coverage. Information, Communication & Society, 15, 3 (2012), 394-410.
[12] Stephen Brammer, Layla Branicki and Martina K Linnenluecke. 2020. COVID-19, Societalization, and the Future of Business in Society. Academy of Management Perspectives, 34, 4 (2020), 493-507.
[13] David Bretthauer. 2001. Open source software: A history. Published Works. 7. (2001).
[14] Maximilian Capraro and Dirk Riehle. 2016. Inner source definition, benefits, and challenges. ACM Computing Surveys (CSUR), 49, 4 (2016), 1-36.
[15] Kenneth E Carpenter. 1972. European industrial exhibitions before 1851 and their publications. Technology and Culture, 13, 3 (1972), 465-486.
[16] Paul E Ceruzzi. 2003. A history of modern computing. MIT press, Cambridge and London.
[17] Henry W Chesbrough. 2003. Open innovation: The new imperative for creating and profiting from technology. Harvard Business Press, Boston, Massachusetts.
[18] Henry W Chesbrough and Melissa M Appleyard. 2007. Open Innovation and Strategy. California Management Review, 50, 1 (2007), 57-77.
[19] Oliver Clarke. 1987. Industrial Democracy in Great Britain. International Studies of Management & Organization, 17, 2 (1987), 38-51.
[20] Diana Crane. 1972. Invisible colleges; diffusion of knowledge in scientific communities. University of Chicago Press, Chicago, Illinois.
[21] Paul A David. 2004. Understanding the emergence of 'open science'institutions: functionalist economics in historical context. Industrial and Corporate Change, 13, 4 (2004), 571-589.
[22] Paul A David. 2005. From keeping 'nature's secrets' to the institutionalization of 'open science'. CODE: Collaborative Ownership and the Digital Economy (2005), 85-108.
[23] Paul A David. 2008. The Historical Origins of 'Open Science': an essay on patronage, reputation and common agency contracting in the scientific revolution. Capitalism and Society, 3, 2, Article 5 (2008).
[24] Phyllis M Deane. 1979. The first industrial revolution. Cambridge University Press, Cambridge, UK.
[25] Laura Dobusch, Leonhard Dobusch and Gordon Müller-Seitz. 2019. Closing for the benefit of openness? The case of Wikimedia's open strategy process. Organization Studies, 40, 3 (2019), 343-370.
[26] Arthur L Dunham. 1955. The industrial revolution in France, 1815-1848. Exposition Press, New York, New York.
[27] Walter G Endrei. 1968. The first technical exhibition. Technology and Culture, 9, 2 (1968), 181-183.
[28] Brian Fitzgerald. 2006. The transformation of open source software. MIS Quarterly, 30, 3 (2006), 587-598.
[29] Julien Foudraine. 2015. Practices to Involve Employees in the Strategy Process. In Proceedings of the IBA Bachelor Thesis Conference, University of Twente (Enschede, Netherlands, 2015).
[30] Chris Freeman. 2019. History, co-evolution and economic growth. Industrial and Corporate Change, 28, 1 (2019), 1-44.
[31] Daniel D Garcia-Swartz and Martin Campbell-Kelly. 2019. Openness as a business strategy: Historical perspectives on openness in computing and mobile phones. Information Economics and Policy, 48 (2019), 1-14.
[32] Michael J Gill, David J Gill and Thomas J Roulet. 2018. Constructing trustworthy historical narratives: Criteria, principles and techniques. British Journal of Management, 29, 1 (2018), 191-205.
[33] Henry P Guzda. 1984. Industrial democracy: made in the USA. Monthly Labor Review, 107 (1984), 26-33.
[34] Daryl M Hafter. 1984. The business of invention in the Paris Industrial Exposition of 1806. Business History Review (1984), 317-335.
[35] Chris Hart. 2018. Doing a literature review: Releasing the research imagination. SAGE, London, UK.
[36] Julia Hautz, Katja Hutter, Johannes Sutter and Johannes Füller 2019. Practices of inclusion in open strategy. In: D. Seidl, G. von Krogh and R. Whittington (eds.) Cambridge Handbook of Open Strategy. Cambridge University Press, Cambridge, UK.
[37] Julia Hautz, David Seidl and Richard Whittington. 2017. Open strategy: Dimensions, dilemmas, dynamics. Long Range Planning, 50, 3 (2017), 298-309.
[38] Joachim Hemer. 2011. A snapshot on crowdfunding. Arbeitspapiere Unternehmen und Region, No. R2/2011, Fraunhofer ISI, Karlsruhe, 2011.
[39] Eric von Hippel and Georg von Krogh. 2003. Open source software and the "private-collective" innovation model: Issues for organization science. Organization Science, 14, 2 (2003), 209-223.
[40] Eric Hopkins. 1982. Working hours and conditions during the Industrial Revolution: A re-appraisal. Economic History Review (1982), 52-66.
[41] Sara Horrell and Jane Humphries. 1995. "The exploitation of little children": Child labor and the family economy in the industrial revolution. Explorations in Economic History, 32, 4 (1995), 485-516.
[42] Mokter Hossain and KM Zahidul Islam. 2015. Ideation through online open innovation platform: Dell IdeaStorm. Journal of the Knowledge Economy, 6, 3 (2015), 611-624.
[43] Jeff Howe. 2006. The rise of crowdsourcing. Wired Magazine, 14, 6 (2006), 1-4.
[44] Noor Huijboom and Tijs Van den Broek. 2011. Open data: an international comparison of strategies. European Journal of ePractice, 12, 1 (2011), 4-16.
[45] Katja Hutter, Bright Adu Nketia and Johann Füller. 2017. Falling Short with Participation — Different Effects of Ideation, Commenting, and Evaluating Behavior on Open Strategizing. Long Range Planning, 50, 3 (2017), 355-370.
[46] Richard Hyman. 2016. The very idea of democracy at work. Transfer: European Review of Labour and Research, 22, 1 (2016), 11-24.
[47] Marijn Janssen, Yannis Charalabidis and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29, 4 (2012), 258-268.

[48] Henry F Kaiser. 1960. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20, 1 (1960), 141-151.
[49] Morton Keller. 1997. The making of the modern corporation. The Wilson Quarterly, 21, 4 (1997), 58-69.
[50] Martin Kenney and John Zysman. 2016. The rise of the platform economy. Issues in Science and Technology, 32, 3 (2016), 61-69.
[51] B Zorina Khan. 2015. Inventing prizes: a historical perspective on innovation awards and technology policy. Business History Review, 89, 4 (2015), 631-660.
[52] Barbara Kitchenham and Stuart Charters. 2007. Guidelines for Performing Systematic Literature Reviews in Software Engineering. Keele University & University of Durham, 2007.
[53] David S Landes. 2003. The unbound Prometheus: technological change and industrial development in Western Europe from 1750 to the present. Cambridge University Press, Cambridge, UK.
[54] Terence H Lloyd. 2002. England and the German Hanse, 1157-1611: a study of their trade and commercial diplomacy. Cambridge University Press, Cambridge, UK.
[55] James H Love, Stephen Roper and Priit Vahter. 2014. Learning from openness: The dynamics of breadth in external innovation linkages. Strategic Management Journal, 35, 11 (2014), 1703-1716.
[56] William Mass. 1989. Mechanical and organizational innovation: The Drapers and the automatic loom. Business History Review (1989), 876-929.
[57] Kurt Matzler, Johann Füller, Katja Hutter, Julia Hautz and Daniel Stieger. 2014. Social Media and Open Strategy: Towards a Research Agenda. In Proceedings of the 22nd European Conference on Information Systems (Tel Aviv, Israel, 2014).
[58] Richard A May. 1923. The Trade Association and its Place in the Business Fabric. Harvard Business Review, 2, 1 (1923), 84-97.
[59] Petra Moser. 2013. Patents and innovation: evidence from economic history. Journal of Economic Perspectives, 27, 1 (2013), 23-44.
[60] Thomas R Navin. 1950. The Whitin Machine Works since 1831: A Textile Machinery Company in an Industrial Village. Harvard University Press, Boston, Massachusetts.
[61] Sheilagh Ogilvie. 2011. Institutions and European trade: Merchant guilds, 1000–1800. Cambridge University Press, Cambridge, UK.
[62] Bruce Perens. 1999. The open source definition. Open sources: voices from the open source revolution, 1 (1999), 171-188.
[63] Sandra Peter and Markus Deimann. 2013. On the role of openness in education: A historical reconstruction. Open Praxis, 5, 1 (2013), 7-14.
[64] HH Rachford Jr and JD Rice. 1952. Procedure for use of electronic digital computers in calculating flash vaporization hydrocarbon equilibrium. Journal of Petroleum Technology, 4, 10 (1952), 19.
[65] Maximilian Rapp, Markus Rhomberg, Giordano Koch and Ken White. 2016. A New Path for the Public Sector: How to Design a Co-created Strategy in Higher Education. In Proceedings of the International Conference on Electronic Participation (Guimarães, Portugal, 2016).
[66] Eric Raymond. 1999. The cathedral and the bazaar. Knowledge, Technology & Policy, 12, 3 (1999), 23-49.
[67] Roy Rosenzweig. 2006. Can history be open source? Wikipedia and the future of the past. The Journal of American History, 93, 1 (2006), 117-146.
[68] Michael Rothmann. 1998. Die Frankfurter Messen im Mittelalter. Franz Steiner Verlag, Stuttgart, Germany.
[69] Daniel Schlagwein, Kieran Conboy, Joseph Feller, Lorraine Morgan and Jan Marco Leimeister. 2017. 'Openness' With and Without IT: A Framework and a Brief History. Journal of Information Technology, 32, 4 (2017), 297-305.
[70] Suntje Schmidt and Verena Brinks. 2017. Open creative labs: Spatial settings at the intersection of communities and organizations. Creativity and Innovation Management, 26, 3 (2017), 291-299.
[71] David Seidl and Felix Werle. 2018. Inter-organizational sensemaking in the face of strategic meta-problems: Requisite variety and dynamics of participation. Strategic Management Journal, 39, 3 (2018), 830-858.
[72] Julian A Smith. 1992. Precursors to Peregrinus: The early history of magnetism and the mariner's compass in Europe. Journal of Medieval History, 18, 1 (1992), 21-74.
[73] Matthew L Smith and Ruhiya Seward. 2017. Openness as social praxis. First Monday (2017).
[74] Peter HA Sneath. 1957. The application of computers to taxonomy. Microbiology, 17, 1 (1957), 201-226.
[75] Hemang C Subramanian and Suresh Malladi. 2020. Bug Bounty Marketplaces and Enabling Responsible Vulnerability Disclosure: An Empirical Analysis. Journal of Database Management (JDM), 31, 1 (2020), 38-63.
[76] Asin Tavakoli, Daniel Schlagwein and Detlef Schoder. 2017. Open strategy: Literature review, re-analysis of cases and conceptualisation as a practice. The Journal of Strategic Information Systems, 26, 3 (2017), 163-184.
[77] Ross E Traub, Joel Weiss, CW Fisher and Don Musella. 1972. Closure on openness: Describing and quantifying open education. Interchange, 3, 2-3 (1972), 69-84.
[78] Ilkka Tuomi. 2001. Internet, innovation, and open source. First Monday (2001).
[79] Rubén Vicente-Sáez and Clara Martínez-Fuentes. 2018. Open Science now: A systematic literature review for an integrated definition. Journal of Business Research, 88 (2018), 428-436.
[80] Blair Wang, Daniel Schlagwein, Dubravka Cecez-Kecmanovic and Michael C Cahalane. 2018. Digital work and high-tech wanderers: Three theoretical framings and a research agenda for digital nomadism. In Proceedings of the Australasian Conference on Information Systems (Sidney, Australia, 2018).
[81] Sidney Webb and Beatrice Webb. 1897. Industrial democracy. Longmans, Green & Co., London, New York and Bombay.
[82] Sidney Webb and Beatrice Webb. 1920. The history of trade unionism. Longmans, Green & Co., London, New York, Bombay, Calcutta, Madras.
[83] Steven Weber. 2004. The Success of Open Source. Harvard University Press, Boston, Massachusetts.
[84] Richard Whittington. 2015. The Massification of Strategy. British Journal of Management, 26 (2015), 13-16.
[85] Richard Whittington, Ludovic Cailluet and Basak Yakis-Douglas. 2011. Opening Strategy: Evolution of a Precarious Profession. British Journal of Management, 22, 3 (2011), 531-544.
[86] Michael Widenius, David Axmark and Kaj Arno. 2002. MySQL reference manual: documentation from the source. O'Reilly Media, Inc., Sebastobol, California.
[87] Brian Winston. 2002. Media, technology and society: A history: From the telegraph to the Internet. Routledge, London and New York.

Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression

Nathan TeBlunthuis
nathante@uw.edu
University of Washington
Seattle, Washington, USA

ABSTRACT

Organizing complex peer production projects and advancing scientific knowledge of open collaboration each depend on the ability to measure quality. Wikipedia community members and academic researchers have used article quality ratings for purposes like tracking knowledge gaps and studying how political polarization shapes collaboration. Even so, measuring quality presents many methodological challenges. The most widely used systems use quality assesements on discrete ordinal scales, but such labels can be inconvenient for statistics and machine learning. Prior work handles this by assuming that different levels of quality are "evenly spaced" from one another. This assumption runs counter to intuitions about degrees of effort needed to raise Wikipedia articles to different quality levels. I describe a technique extending the Wikimedia Foundations' ORES article quality model to address these limitations. My method uses weighted ordinal regression models to construct one-dimensional continuous measures of quality. While scores from my technique and from prior approaches are correlated, my approach improves accuracy for research datasets and provides evidence that the "evenly spaced" assumption is unfounded in practice on English Wikipedia. I conclude with recommendations for using quality scores in future research and include the full code, data, and models.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**; *Social content sharing*; **Computer supported cooperative work**.

KEYWORDS

sociotechnical systems, measurement, statistics, quality, machine learning, peer production, Wikipedia, online communities, methods, datasets

ACM Reference Format:
Nathan TeBlunthuis. 2021. Measuring Wikipedia Article Quality in One Dimension by Extending ORES with Ordinal Regression. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3479986.3479991

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479991

1 INTRODUCTION

Measuring content quality in peer production projects like Wikipedia is important so projects can learn about themselves and track progress. Measuring quality also helps build confidence that information is accurate and supports monitoring how well an encyclopedia includes diverse subject areas to identify gaps needing attention [31]. Measuring quality enables tracking and evaluating the progress of subprojects and initiatives organized to fill the gaps [16, 41]. Raising an article to a high standard of quality is a recognized achievement among contributors, so assessing quality can help motivate contributions [5, 14]. In these ways, measuring quality can be of key importance to advancing the priorities of the Wikimedia movement and is also important to other kinds of open collaboration [10].

Measuring quality also presents methodological and ontological challenges. How can "quality" be conceptualized so that measurement of the goals of a project and the value it produces can be precise and accurate? Language editions of Wikipedia, including English, peer produce quality labels that have been useful both for motivating and coordinating project work and for enabling research. Epistemic virtues of this approach stem from the community-constructed criteria for assessment and from formalized procedures for third-party evaluation organized by WikiProjects. These systems also have two important limitations: (1) ratings are likely to lag behind changes in article quality, and (2) quality is assessed on a discrete ordinal scale, which violates typical assumptions in statistical analysis. Both limitations are surmountable.

The machine learning framework introduced by Warncke-Wang et al. [42], further developed by Halfaker [16], implemented by the Objective Revision Evaluation Service[1] (ORES) article quality models and adopted by several research studies of Wikipedia article quality [e.g. 17, 22, 34, 41] was designed to address the first limitation by using article assessments at the time they were made as "ground truth." Article quality might drift in the periods between assessments, but it seems safe to assume that new quality assessments are accurate at the time they are made. A model trained on recent assessments can predict what quality label an article would receive if assessed in its current state.

This paper introduces a method for constructing interpretable one-dimensional measures of article quality from Wikipedia quality assessments and the ORES article quality model. The method improves upon prior approaches in two important ways. First, by using inverse probability weighting to calibrate the model, it is more accurate for typical research applications, and second, it does not depend on the assumption that quality levels are "evenly spaced," which threatens the validity of prior research [4, 16]. In addition,

[1] https://www.mediawiki.org/wiki/ORES (https://perma.cc/TH6L-KFT6)

this paper helps us understand the validity of previous work by analyzing the performance of the ORES quality model and testing the "evenly spaced" assumption.

In §2, I provide a brief overview of quality measurement in peer production research in which I foreground the importance of the assumptions needed to use machine learning predictions in downstream analysis—particularly the "evenly spaced" assumption used by Halfaker [16] to justify the use of a handpicked weighted sum to combine article class probabilities. Next, in §3, I describe how to build accurate ordinal quality models that are appropriately calibrated for analyses of representative samples of Wikipedia articles or revisions. I also briefly explain how ordinal regression provides an interpretable one-dimensional measure of quality and how it relaxes the "evenly spaced" assumption. Finally, in §4 I present the results of my analysis to (1) show how the precision of the measurement depends on proper calibration and (2) demonstrate that the "evenly spaced" assumption is violated. Despite this, I find that scores from the ordinal models are highly correlated with those from prior work so the "evenly spaced" assumption may be acceptable in some applications. I conclude in §5 with recommendations for measuring article quality in future research.

2 BACKGROUND

Measurement is important to science as available knowledge often constrains the development of improved tools for advancing knowledge. For example, in the book *Inventing Temperature*, Hasok Chang [11], the philosopher and historian of science, documents how extending theories of heat beyond the range of human sense perception required scientists to develop new types of thermometers. This in turn required better knowledge of heat and of thermometric materials such as the freezing point of mercury. Part of the challenge of scientific advancement is that measurement devices developed under certain conditions may give unexpected results outside of the range in which they are calibrated: a thermometer will give impossibly low temperature readings when its mercury unexpectedly freezes. Today, machine learning models are used to extend the range of quality measurements in peer production research, but state of the art machine learning can be quite sensitive to the nuances of how their training data are selected [30].

2.1 Measuring Quality in Peer Production

As described in §1, measuring quality has been of great importance to peer production projects like Wikipedia and in the construction of knowledge about how such projects work. The foundation of article quality measurement in Wikipedia has been the peer production of article quality assessment organized by WikiProjects who develop criteria for articles in their domain [28]. This enables quality assessment to be consistent across different subject areas, but the procedures for assessing quality are tailored to the values of each WikiProject. Yet, like human sense perception of temperature, these quality assessments are limited in that they require human time and attention. In addition, humans' limited ability to discriminate between levels on a scale limits the sensitivity of quality assessments. Articles are assessed irregularly and infrequently at the discretion of volunteer editors. Therefore, for most article revisions, it is not known what quality class the article would be assigned if it were newly assessed.

Researchers have proposed many ideas to extend the range of quality measurement beyond the direct perception of Wikipedians, such as page length [7], persistent word revisions [1, 6], collaboration network structures [29], and template-based flaw detection [3]. Carefully constructed indexes benchmarked against English language Wikipedia quality assessments might allow quality measurement of articles that have not been assessed or in projects that have underproduced article assessments [24]. However, such indexes may lack emic validity if they fail to capture important aspects of quality or if notions of quality vary between linguistic communities and might even shape the editing activity in unexpected ways that could ultimately defeat their purpose [15, 35]. Peer-produced quality labels depend on the limited capacity of volunteer communities to coordinate quality assessment, but also provide impressive validity for evaluating projects on their own terms.

2.2 Article Quality Models Extend Measurement to Unassessed Articles

Perhaps the most successful approaches to extending the range of quality measurements use machine learning models trained on available article quality assessments to predict the quality of revisions that have not been assessed. The ORES article quality model (henceforth ORES) implements this approach, but other similar article quality predictors have been developed [2, 12, 13, 29, 32, 44], and additional features including those based on language models can substantially improve classification performance compared to ORES [33]. The ORES model is a tree-based classifier that predicts the quality class of a Wikipedia article at the time it is assessed.[2] These tree-based models are reasonable for practical purposes with the reported ability to predict within one level of the true quality class with 90% accuracy (although in §4.2 I find a decline in accuracy in a more recent dataset). Yet, since these models do not account for the ordering of quality labels, the use of these predictions in downstream analysis introduces complicated methodological challenges.

The ORES classifiers are fit using `scikit-learn`[3] through minimization of the multinomial deviance as shown [18, 27]:

$$L(y_i, p(x_i)) = -\sum_{k=1}^{K} I(y_i = \mathcal{G}_{i,k}) \log p_k(x_i) \tag{1}$$

For each article i with predictors x_i that has been labeled with a quality class y_i, the ORES model outputs an estimated probability $p_k(x_i)$ that the article belongs to each quality class $k \in \{\textit{stub}, \textit{start}, \textit{C-class}, \textit{B-class}, \textit{Good article (GA)}, \textit{Featured article (FA)}\}$. The predicted probabilities $p(x_i)$ sum to one so the ORES model outputs a unit vector for each article. If $\mathcal{G}_{i,k}$, the most probable quality class (MPQC) according to the model, is the true label, then $I(y_i = \mathcal{G}_{i,k})$ equals 1 (I is the indicator function) and the log predicted probability $p_k(x_i)$ of the correct class is subtracted from the loss $L(y_i, p(x_i))$. Note that this model does not use the fact that

[2] The system uses cross-validation to select among candidates that include random-forest and boosted decision tree models.

[3] https://scikit-learn.org/stable/(https://perma.cc/5Y8B-W8T5)

article quality classes are ordered. If it did, then it would have to penalize an incorrect classification of a *Good article* as *C-class* more than a classification of a *Good article* as *B-class*. In this model, different quality classes have no intrinsic rank or ordering and thus are akin to different categories of article subjects like animals, vegetables, or minerals.

The MPQC is perhaps the most natural way to use the ORES output to measure quality. It has been used in several studies including to provide evidence that politically polarized collaboration on Wikipedia leads to high quality articles [34] and to understand the relationship between article quality and donation [22]. However, the MPQC is limited in that it does not measure quality differences between articles that have the same MPQC. Consider two hypothetical articles; the first has the multinomial prediction $(0.1, 0.3, 0.4, 0.075, 0.075, 0)$ and the second has the prediction $(0.075, 0.075, 0.4, 0.3, 0.1, 0)$. The MPQC will assign both the *C-class* label even though the first article has the same chance at being a *Stub* or *Start-class* as the second article's chance at being a *B-class* or even a *Good article*. At best, the MPQC has limited sensitivity to subtle variations or gradual changes in quality [16].

2.3 Combining Scores for Granular Measurement

To further extend the range of article quality measurement within article quality classes, Halfaker [16] constructed a numerical quality score using a linear combination (a weighted sum) of the elements of the multinomial prediction $p(x_i)$. This is advantageous from a statistical perspective as it naturally provides a continuous measure of quality which can typically justify a normal or log-normal statistical model. It can also support higher-order aggregations for measuring the quality of a set of articles [16]. Halfaker handpicks the coefficients $[0, 1, 2, 3, 4, 5]$ to make a linear combination of the predictions under the assumption "that the ordinal quality scale developed by Wikipedia editors is roughly cardinal and evenly spaced," which I refer to the "evenly spaced" assumption. It essentially says that a *Start-class* article has one more unit quality of a *Stub-class* article, and that a *C-class* article has one more unit of quality than a *Start-class* article and so on. This approach is being adopted by other researchers including Arazy et al. [4].

The considerable degree of effort and expertise required to raise articles to higher levels of quality raises doubt in the assumption [20]. Higher quality levels correspond to increasing completeness, encyclopedic character, usefulness to wider audiences, incorporation of multimedia, polished citations, and adherence to Wikipedia's policies. The English language Wikipedia editing guideline on content assessment[4] defines a *Good article* as "useful to nearly all readers, with no obvious problems" and a *Featured article* article as "professional, outstanding and thorough." According to Wikipedians, it can take "three to six months of full time work" to write a *Featured article*.[5] Are we to assume that the difference in quality between a *Good article* and a *Featured article* is measurably the same as that between a *Stub* defined as as "little more than a dictionary definition" and a *Start-class* that is "a very basic description of the topic?" How could we even answer this question?

If the "evenly spaced" assumption is reasonable, then Halfaker's [16] weighted sum approach is too. But if increasing Wikipedia article classes do not represent roughly equal improvements in quality, this may threaten the accuracy of analysis dependent on the assumption. Suppose that a *B-class* article has not 1, but 2 units of quality greater than a *C-class* article, then Halfaker could have underestimated the improvement in the knowledge gap of women scientists, which was considerably driven by improvement in *B-class* articles. In the next section, I provide a straightforward extension of the ORES article quality model based on ordinal regression that both relaxes the "evenly spaced" assumption and provides a better calibrated and more accurate one-dimensional measure of quality.

3 DATA, METHODS AND MEASURES

I use Bayesian ordinal regression models that use the ORES predicted probabilities to predict the quality class labels and quantify the distance between quality classes. I now provide a brief overview of ordinal regression as needed to explain my approach to measuring quality. Understanding ordinal regression depends on background knowledge of odds and generalized linear models. I recommend McElreath and Safari [25] for reference.

3.1 Bayesian Ordinal Regression

Ordinal regression predicts quality class membership using a single linear model for all classes and identifies boundaries between classes using the log cumulative odds link function shown below in Eq. 2. The log cumulative odds is not the only possible choice of link function, but it is the most common, is the easiest to interpret, and is appropriate here.

$$\begin{aligned} &\log \frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)} = \alpha_k - \phi_i \\ &\phi_i = Bx_i \end{aligned} \tag{2}$$

As in Eq. 1, y_i is the quality label for article i. The left hand side of Eq. 2 gives the log odds that y_i is less than or equal to quality level k. The ordinal quality measure is given by a linear model $\phi_i = Bx_i$ (x_i is a vector of transformed ORES scores for article i). Key to interpreting ϕ_i as a quality measure are the intercept parameters a_k for each quality level k. The log cumulative odds (the log odds that the article y_i has quality less than or equal to k) are given by the difference between the intercept and the linear model a_k - ϕ_i. Therefore, if $\phi_i = \alpha_k$ then the chances that $i <= k$ equal the chances that $i > k$. When ϕ_i is less than α_k, the quality of article i is probably less than or equal to quality level k. As $\phi_i - \alpha_k$ increases so do the chances that article i is of quality better than k. In this way, the threshold parameters a_k define quantitative article quality levels on the scale of the ordinal quality measure ϕ_i.

Informally, an ordinal regression model maps a linear regression model to the ordinal scale using the log cumulative odds link function. It does this by inferring thresholds that partition the range of linear predictions. When the linear predictor for an article crosses a threshold, the probability that the article has quality greater than that corresponding to the threshold begins to increase.

[4] https://en.wikipedia.org/w/index.php?title=Wikipedia:Content_assessment&oldid=1023695750 (https://perma.cc/2JUV-6SD)

[5] Public statement by Stuart Yeates, an expert Wikipedian; quoted with permission. https://lists.wikimedia.org/hyperkitty/list/wiki-research-l@lists.wikimedia.org/message/7U35LHAXRWEPABN75DOTPOIEA2VYCTQQ/ (https://perma.cc/9V4P-WRXR)

Table 1: Numbers of articles and revisions, sample sizes, and regression weights for each quality level.

Label	No. of articles	No. of revisions	Sample size	Article weights	Revision weights
Stub	3,359,351	12,005,611	4,969	4.23	2.52
Start	1,019,038	7,828,335	4,979	1.28	1.64
C	235,655	3,889,639	4,988	0.30	0.81
B	128,875	3,640,591	4,990	0.16	0.76
GA	31,808	924,468	4,999	0.04	0.19
FA	7,438	365,255	4,995	0.01	0.08

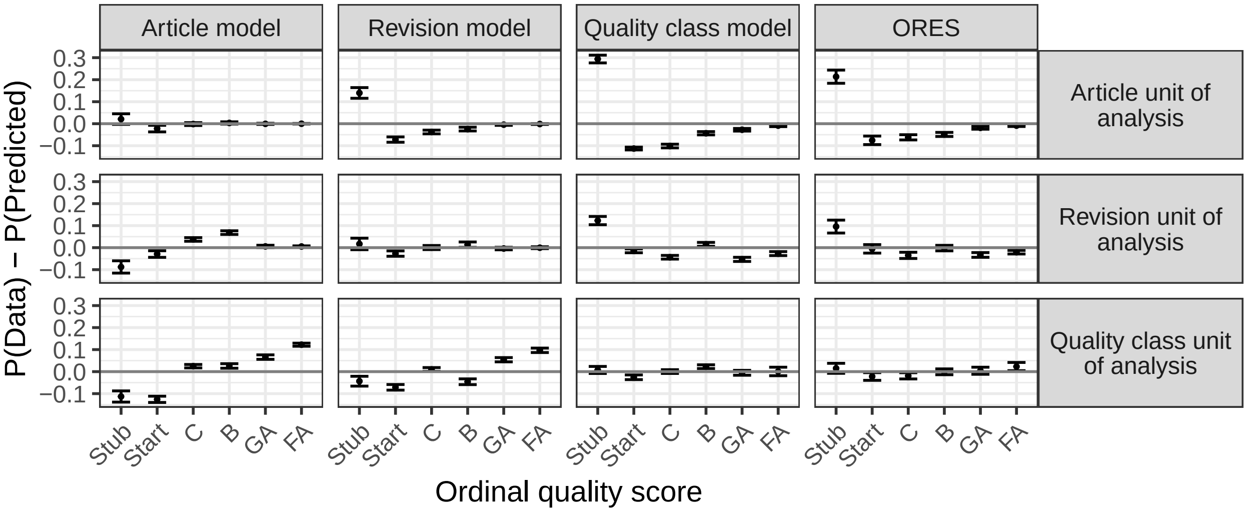


Figure 1: Calibration of each predictive quality model on datasets representative of each unit of analysis (article, revision, quality class). Each chart shows, for each quality class, the miscalibration of a model (columns) with respect to a dataset weighted to represent a unit of analysis (rows). The y-axis shows difference between the true probability of the quality class and the average predicted probability of that class, given a chosen unit of analysis. Points close to zero indicate good calibration. For example, the top-left chart shows that the article model is well-calibrated to the dataset on which it was fit and the middle-left chart shows that the article model predicts that articles are *Stubs* with probability greater than the frequency of *Stubs* in a random sample of revisions. Error bars show 95% confidence intervals.

Bayesian inference allows interpreting model parameters like ϕ_i and α_k as random variables and provides accurate quantification of uncertainty in thresholds and predictions. I fit models using the R package Bayesian regression modeling using Stan (`brms`) [8] version 2.15.0. I use the default priors for ordinal regression, which are weakly informative. Due to the large sample size, the data overwhelm the priors and the priors have little influence over results. I confirmed this by fitting equivalent frequentist models using the `polr` function in the `MASS` R package [40] and found that the estimates of intercepts and coefficients were very close.

The six quality scores output by the ORES article quality classifier are perfectly collinear by construction because they sum to one. This means they cannot all be included in the same regression model. Since interpreting the coefficients is not important, I take the linear transformation of the ORES scores using appropriately weighted principle component analysis and use the first five principle components as the independent variables. This is simpler and more statistically efficient than a model selection procedure.

3.2 Dataset and Model Calibration

I draw a new random sample of 5,000 articles from each quality class to develop my models. I first reuse code from the `articlequality`[6] Python package to process the March 2020 XML dumps for English Wikipedia and extract up-to-date article quality labels. I then select pages that have been assessed by a member of at least one WikiProject. Following prior work, if an article is assessed at different levels according to more than one WikiProject, I assign it to the highest such level and I drop articles having the rarely used *A-class* quality level [16, 41, 42]. Next, I use the `revscoring`[7] Python package to obtain the ORES scores of the labeled article versions. Some of these versions have been deleted leading to missing observations at each quality level. Table 1 shows the number of articles sampled in each quality class. I reserve a random sample of 2000 articles which I use in reporting my results and fit my ordinal regression models on the remainder.

[6] https://pypi.org/project/articlequality (https://perma.cc/8R4H-MAZ9)
[7] https://pypi.org/project/revscoring (https://perma.cc/3HFN-V23Z)

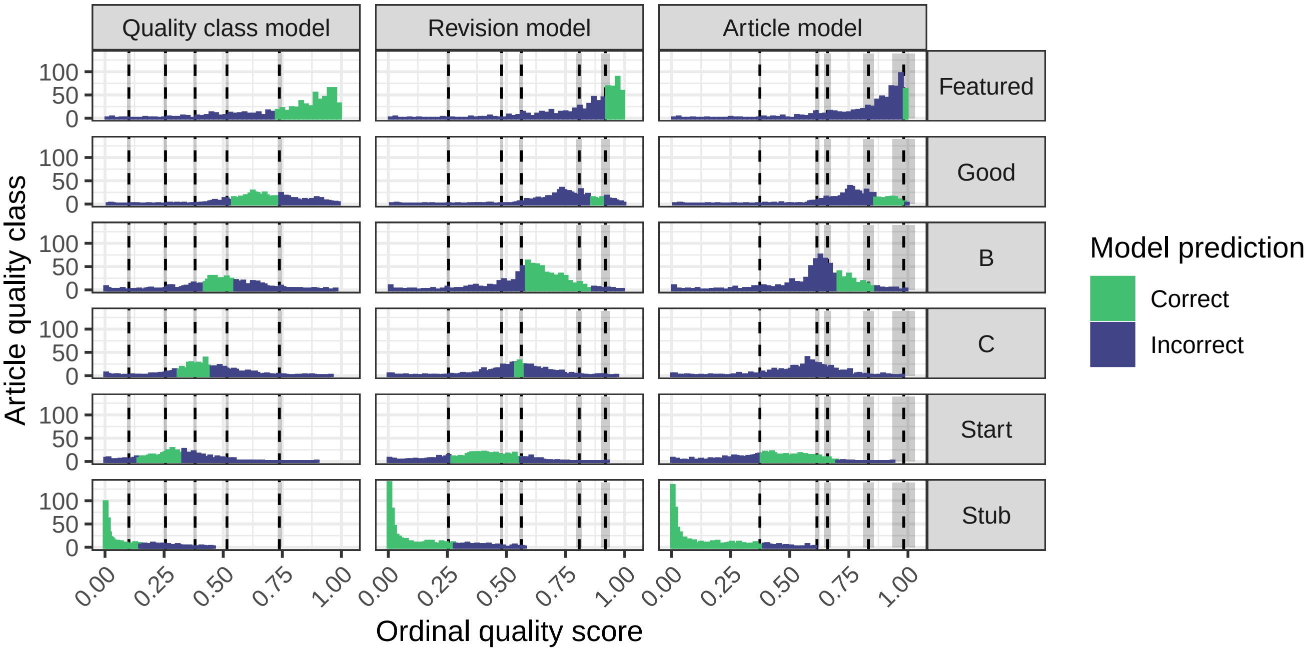


Figure 2: Quality scores and predictions of the ordinal regression models. Columns in the grid of charts correspond to the ordinal quality model calibrated to the indicated unit of analysis and rows correspond to sampled articles having the indicated level of quality as assessed by Wikipedians. Each chart shows the histogram of scores, thresholds inferred by the ordinal model with 95% credible intervals colored in gray, and colors indicating when the model makes correct or incorrect predictions. The thresholds are not evenly spaced, especially in *revision model* and *article model* that has more weight on lower quality classes. These two models infer that the gaps between *Stub* and *Start* and between *Start* and *C-class* articles are considerably wider than the gap between *C-class* and *B-class* articles.

The ORES article quality classifiers are fit on a "balanced" dataset having an equal number of articles in each quality class. Thus, an ORES score is the probability that an article is a member of a quality class under the assumption that the article was drawn from a population where each quality class contains an equal number of articles. Simply put, the model has learned from its training data that each quality class is about the same size.

This is not representative of the overall article quality on Wikipedia, which is highly skewed with over 3 million *Stubs* but only around *7,000 Featured articles* as shown in Table 1. Although using a balanced dataset likely improves the accuracy of the ORES models, for the ordinal regression models, the choice of unit of analysis presents a trade-off between accuracy in a representative sample of articles or revisions and accuracy within each quality class. Constructing a balanced dataset by oversampling is a common practice in machine learning because it can improve predictive performance. However, oversampling can also lead to badly calibrated predictive probabilities as shown in Fig. 1. Calibration means that, on average, the predicted probability of a quality class equals the average true probability of that class for the unit of analysis.

The "balanced" dataset on which ORES is trained has the *quality class* unit of analysis because each quality class has equal representation. However, researchers are more interested in analyzing representative samples of *articles* or *revisions*. For example, the article unit of analysis would be used to estimate the average quality of a random sample of articles and the revision unit of analysis might be used to model the change in the quality of an encyclopedia over time. Weighting allows the use of the balanced dataset to estimate a model as if the dataset were a uniform random sample of a different unit of analysis. My method uses a balanced dataset to fit ordinal regression models with inverse probability weighting to calibrate each model to the unit of analysis of a research project. For example, each article in the model calibrated to the article unit of analysis is weighted by the probability of its quality class in the population of articles divided by the probability of its quality class in the sample. The size of the sample and the weights for the article and revision levels of analysis are also shown in Table 1.

4 RESULTS

I first report my findings about the spacing of the quality classes in each of the models in §4.1. Quality classes are not evenly spaced, especially when articles or revisions are the unit of analysis. Next, in §4.2, I report the accuracy of each of the models and the uncertainty of the ordinal quality scale. All models perform similarly to or better than the MPQC within the pertinent unit of analysis. The unweighted model provides the best accuracy and lowest uncertainty across the entire range of quality levels, but is poorly

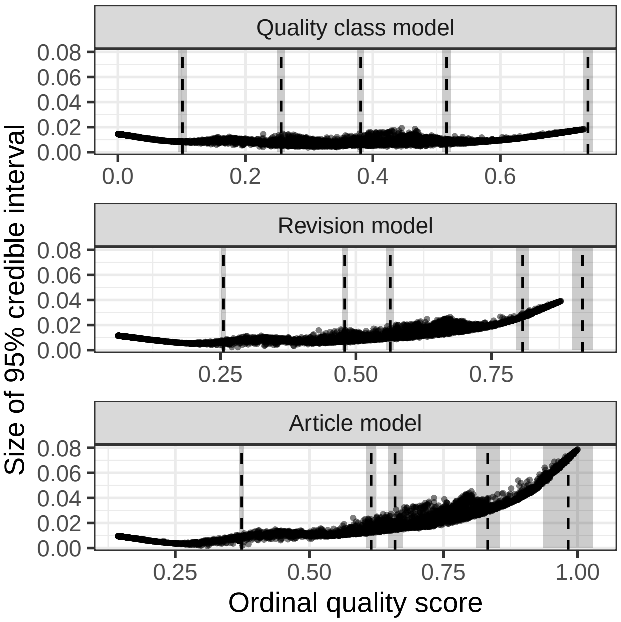


Figure 3: Uncertainty in ordinal quality scores for models calibrated at each unit of analysis. Points show the size of the 95% credible interval for the ordinal quality score for each article in the dataset. The quality class model has low uncertainty across the range of quality. Models calibrated to the revision and article levels of analysis have less uncertainty at the low end of the quality scale, but greater uncertainty at the higher end of the scale.

calibrated for other units of analysis. Finally, in §4.3, I show that all quality measures are highly correlated, but the ordinal quality measures agree with one another more than with the "evenly spaced" measure.

4.1 Spacing of Quality Classes

The grid of charts in Fig. 2 shows quality scores and thresholds for each model (columns) and article quality level (rows). Each chart shows the histogram of quality scores ϕ_i given to articles having the true quality label corresponding to the row of the grid. The histograms are colored to indicate regions where the model correctly predicts that articles belong to their true class. Vertical dashed lines show the thresholds inferred by the model with 95% credible intervals colored in gray. Different models have different ranges of scores, so Fig. 2 shows results normalized between 0 and 1.

No matter the unit of analysis, article quality classes are not evenly spaced. The quality class model provides a quality scale in which *Featured* articles take up 27% of the scale and are expected to score in the range of $[0.73, 1]$, but probable *C-class* articles only span 14% of the scale in the range $[0.31, 0.45]$. Researchers are likely to be interested in models calibrated to the article or revision units of analysis, and in these cases, the quality classes are far from evenly spaced. The *revision model* assigns 28% of the scale to *Stubs*, from 0 to 0.28. It assigns *C-class* articles the smallest part of the scale, only 4% of it, from 0.54 to 0.58. The *article model* is even more extreme. It assigns *Stubs* to the interval $[0, 0.39]$, 39% of the scale, and the space between thresholds defining the range of *C-class* articles is so narrow that it virtually never predicts that an article will be C-class. In general terms, the *quality class model* gives relatively equal amounts of space to each quality class compared to the other models, while reserving nearly the top half of the scale for the top 2 quality classes. The *revision model* and *article model* do the opposite and use the bottom half of the scale to account for differences within the bottom two quality classes, leave some room for *B-class* articles, but squeeze the top end of the scale and *C-class* articles into relatively small intervals.

4.2 Accuracy and Uncertainty

I evaluate predictive performance in terms of *accuracy*, the proportion of predictions of article quality that are correct. To allow comparison with the reported accuracy of the ORES quality models, I also report *off-by-one accuracy*, which includes predictions within one level of the true quality class among correct predictions.

As shown in Table 2, the ordinal regression models have better predictive ability than the MPQC except when the unit of analysis is the quality class. In this case, the best ordinal quality model has worse accuracy than the MPQC but slightly better off-by-one accuracy. Table 2 shows accuracy and off-by-one accuracy weighted for each unit of analysis. Accuracy for a given unit of analysis depends on having a model fit to data representative of that unit of analysis. Accuracy scores are higher when greater weight is placed on lower article quality classes, suggesting that it is easier to discriminate between these classes

The ORES article quality model has been quickly adopted by researchers, but its accuracy is limited. While off-by-one accuracy is above 90% when the article is the unit of analysis, the MPQC only predicts the correct quality class 55% of the time when the quality class is the unit of analysis.

The trade-offs in selecting a unit of analysis on which to calibrate the models are further illustrated by Fig. 3, which plots the size of the 95% credible intervals as a function of the quality scores for each model. As in Fig. 2, quality scores in this plot are rescaled between 0 and 1. The models calibrated to articles or revisions have more certainty in the lower range of the quality scale compared to the model that places equal weight in all quality classes. This comes with a trade-off for the higher range of quality. While the *quality class model* has relatively low uncertainty across the entire range of quality, the *revision model* and *article model* have greater uncertainty at higher levels of quality.

4.3 Correlation Between Scores

Although the models have different predictive performances and uncertainties, as measures of quality, they are nearly perfectly correlated with one another as shown in Fig. 4. For each quality score, including the "evenly spaced" weighted sum, Fig. 4 shows a scatter plot and two correlation statistics: Kendall's τ and Pearson's r. Pearson's r is the standard linear correlation coefficient and Kendall's τ is a nonparametric rank-based correlation defined as the probability that the quality scores will agree about which of

Table 2: Accuracy of quality prediction models depends on the unit of analysis. The greatest accuracy and off-by-one accuracy scores are highlighted. Models are more accurate when calibrated on the same unit of analysis on which they are evaluated. Compared to the MPQC, the ordinal quality models have better accuracy when revisions or articles are the unit of analysis. When the quality class is the unit of analysis, the ordinal quality model has worse accuracy, but predicts within one quality class with slightly better accuracy.

Unit of analysis	Model	Ordinal model?	Accuracy	Off-by-one accuracy
Quality class	Article	Yes	0.33	0.75
Quality class	Revision	Yes	0.44	0.84
Quality class	Quality class	Yes	0.52	0.87
Quality class	ORES MPQC	No	0.55	0.86
Revision	Article	Yes	0.57	0.87
Revision	Revision	Yes	0.61	0.92
Revision	Quality class	Yes	0.54	0.88
Revision	ORES MPQC	No	0.58	0.9
Article	Article	Yes	0.76	0.97
Article	Revision	Yes	0.73	0.96
Article	Quality class	Yes	0.63	0.92
Article	ORES MPQC	No	0.65	0.94

any two articles has higher quality minus the probability that they will disagree.

According to Pearson's r all the quality scores are highly correlated with correlation coefficients of about 0.98 or higher. Kendall's τ measures nonlinear correlation and reveals discrepancies between the ordinal models and the "evenly spaced" measures. The Pearson correlation between scores from the *revision model* and the scores from the *quality class model* are about the same as the correlation between the *revision model* scores and the "evenly spaced" scores ($r = 0.98$). However, according to Kendall's τ, scores from the *revision model* are more similar to those from the *quality class model* ($r = 0.98$) than to the scores from the "evenly spaced" approach ($r = 0.9$).

The evenly spaced model is more likely to disagree with the model-based scores than any of the model-based scores are to disagree with one another as visualized in the scatter plots in Fig. 4. Disagreement between the "evenly spaced" method and the ordinal models is greatest among articles in the middle of the quality range.

5 DISCUSSION

Past efforts to extend the measurement of Wikipedia article quality from peer-produced article quality assessments to unassessed versions of articles and from the discrete to the continuous domain have relied upon machine learning and expedient but untested assumptions like that quality levels are "evenly spaced." While I suggest technical improvements for statistical models for measuring quality, I also find that scores from my models are highly correlated to those obtained under the "evenly spaced" assumption.

I set out to provide a better way to convert the probability vector output by the ORES article quality model into a continuous scale and to test the assumption that the quality levels are evenly spaced. I used ordinal regression models to infer spacing between quality levels and used the linear predictor of these models as a continuous measure of quality. While I found in §4.1 that the quality levels are not evenly spaced and that the spacing depends on the unit of analysis to which the models are calibrated, I also showed in §4.3 that the model-based quality measures are highly, although not perfectly, correlated with the "evenly spaced" measure. This provides some assurance that past results built on this measure are unlikely to mislead. That said, I recommend that future work adopt appropriately calibrated model-based quality measures instead of the "evenly spaced" approach, and I argue that it is important to improve the accuracy of article quality predictors to enable more precise article quality measurement.

5.1 Recommendations for Measuring Article Quality

How should future researchers approach the question of how to measure Wikipedia article quality? While I cannot provide a final or complete answer to the question, I believe the exercise reported in this paper provides some insights on which to base recommendations. It is important to note that I consider here only approaches to measuring quality that assume the use of a good predictor of article quality assessment, such as the ORES quality model. I do not consider other based approaches such as those based on indexes [24] described in §2.

5.1.1 Use the principle components of ORES scores for statistical control of article quality. In many statistical analyses, the only purpose of measuring quality will be as a statistical control or adjustment. For example, Zhang et al. [43] used the MPQC as a control variable in a propensity score matching analysis of promotion to *Featured article* status, but as argued in §3, the MPQC provides less information than the vector of ORES scores. Using the principle components is simpler than using an ordinal quality model. I recommend obtaining ORES scores for your dataset, taking the principle components, and dropping the least significant one to remove collinearity.

5.1.2 Use ordinal quality scores when article quality is an independent variable. In other cases, research questions will ask how article quality is related to an outcome of interest, like how Kocielnik et al. [22] set out to explore factors associated with donations to the

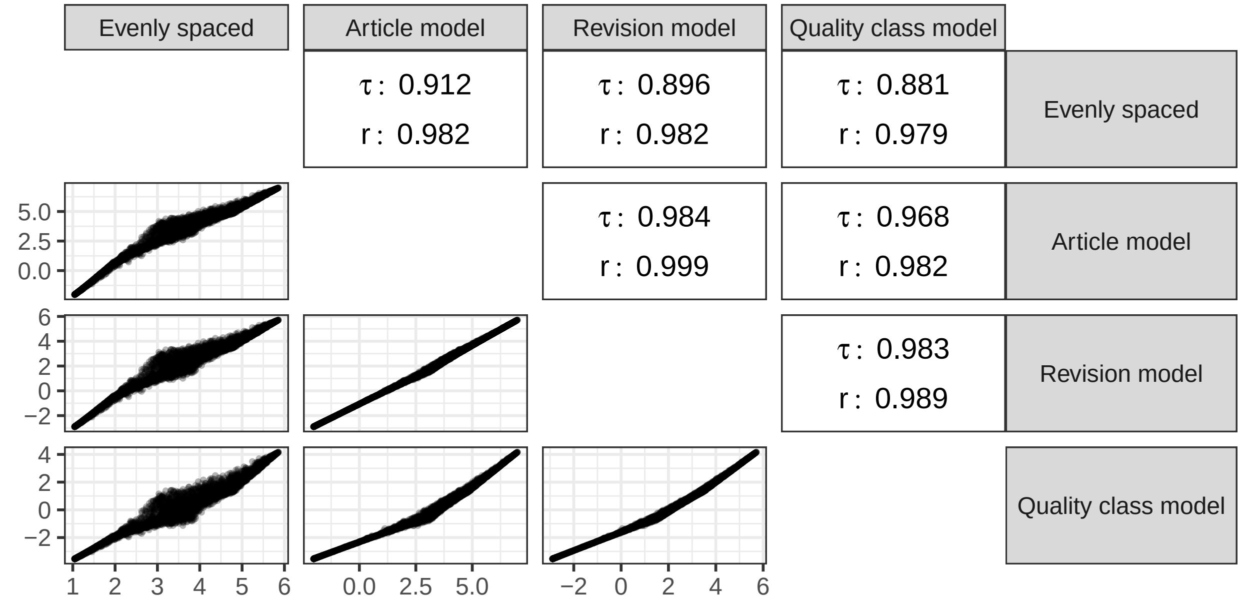


Figure 4: Correlations between quality measures show that the different approaches to measuring quality are quite similar. "Evenly spaced" uses a weighted sum of the ORES scores with handpicked coefficients [16]. Lower values of Kendall's τ, a nonparametric rank correlation statistic, compared to Pearson's r suggest nonlinear differences between the weighted sum and the other measures.

Wikimedia Foundation. They use the MPQC as an independent variable, which complicates their analysis. Although they conclude that "pages with higher quality attract more donations," this is not strictly true. They actually found a nonlinear relationship where readers of *B-class* articles were more likely to donate than readers of *Featured articles.* Using a continuous measure of quality is more convenient when the average linear relationship is the target of inference.

I recommend using an ordinal regression model appropriate to the downstream unit of analysis because this will justify the interpretation of the measure. If the downstream unit of analysis differs substantively from those used here, such as if different selection criteria are applied, I recommend reusing my code to calibrate a new ordinal regression model to a new dataset. Otherwise, reusing one of my models should be adequate. Finally, in the Bayesian framework, the scores are interpretable as random variables. This provides a justification for incorporating the variance of these scores as measurement errors to improve estimation in downstream analysis [25].

5.1.3 Use the MPQC or ordinal quality scores when article quality is the dependent variable. Using the MPQC as the outcome in an ordinal regression model, as is done by Shi et al. [34] in their analysis of Wikipedia articles with politically polarized editors, is a reasonable choice as long as it provides sufficient variation and a more granular quality measure is not needed. Although it is theoretically possible that using the MPQC might introduce statistical bias because it less accurate than ordinal quality scores for units of analysis other than the quality class and omits variation within quality classes, such threats to validity do not seem more significant than the threat introduced by inaccurate predictions. If the MPQC does not provide sufficient granularity and a continuous measure is desired as in Halfaker [16] or Arazy et al. [4], I recommend using a measure based on ordinal regression as described in §5.1.2.

5.2 Limitations

Although intuitions about the varying degrees of effort required to develop articles with different levels of quality led me to question the "evenly spaced" assumption, my findings that quality classes are not evenly spaced do not necessarily reflect relative degrees of effort. Rather, spaces between levels are chosen to link a linear model to ordinal data. The spacing of intervals depends on the ability of the ORES scores to predict quality classes. The ORES article quality model has relative difficulty classifying *C-class* and *B-class* articles [16]. Perhaps, the differences between these quality classes are minor compared to the other classes. Maybe ORES lacks the features or ability to model these differences and the space between these classes will grow if its predictive performance improves.

The usefulness of article quality scores depends on the accuracy of the model. The ORES quality models are accurate enough to be useful for researchers, but they still only predict the correct quality class 55% of the time on a balanced dataset. Of course, this limits the accuracy of the ordinal regression models reported here. Furthermore, while the ORES quality models were designed with carefully chosen features intended to limit biases [17], it is still quite plausible that the accuracy of predictive quality models may vary depending on characteristics of the article [21]. Such inaccuracies may introduce bias, threaten downstream analysis or lead to unanticipated consequences of collaboration tools built upon the models [37]. Therefore, improving the accuracy of article quality prediction models is important to the validity of future article quality research. Adopting machine learning models that can incorporate ordinal loss functions is a promising direction and can reduce the need for auxiliary ordinal regression models [9].

This paper only considers measuring article quality for English language Wikipedia, but expanding knowledge of collaborative encyclopedia production depends on studying other languages as audiences and collaborative dynamics can greatly vary between projects [19, 23, 36]. Other languages carry out quality assessments [24], and some of these have been used to build ORES article quality models. Future work should extend this project to provide multilingual article quality measures in one continuous dimension.

An additional limitation stems from the likelihood that peer-produced quality labels are biased. For instance, the English Wikipedia community has a well-documented pattern of discrimination against content associated with marginalized groups such as biographies of women [26, 38] and indigenous knowledge [39]. Although demonstrating biases in article quality assessment is a task for future research, if Wikipedians' assessments of article quality are biased then model predictions of quality will almost certainly be as well.

6 CONCLUSION

Measuring article quality in one continuous dimension is a valuable tool for studying the peer production of information goods because it provides granularity and is amenable to statistical analysis. Prior approaches extended ORES article quality prediction into a continuous measure under the "evenly spaced" assumption. I showed how to use ordinal regression models to transform the ORES predictions into a continuous measure of quality that is interpretable as a probability distribution over article quality levels, provides an account of its own uncertainty and does not assume that quality levels are "evenly spaced." Calibrating the models to the chosen unit of analysis improves accuracy for research applications. I recommend that future work adopt this approach when article quality is an independent variable in a statistical analysis.

7 CODE AND DATA AVAILABILITY

Code, data and instructions for replicating or reusing this analysis are available in the Harvard Dataverse at https://doi.org/10.7910/DVN/U5V0G1.

ACKNOWLEDGMENTS

I am grateful to the members of the Community Data Science Collective for their feedback on early drafts of this work including Kaylea Champion, Sneha Narayan, Jeremy Foote, and Benjamin Mako Hill. I would also like to thank Aaron Halfaker for encouraging me to write this after seeing a preliminary version. Thanks to Stuart Yeates and other participants in the `wiki-research-l` mailing list wiki-research-l@lists.wikimedia.org for answering my questions about measuring article quality and effort. Finally, thank you to the anonymous OpenSym reviewers whose careful and constructive feedback improved the paper.

REFERENCES

[1] B. Thomas Adler and Luca de Alfaro. 2007. A Content-Driven Reputation System for the Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 261–270.

[2] Maik Anderka and Benno Stein. 2012. A Breakdown of Quality Flaws in Wikipedia. In *Proceedings of the 2Nd Joint WICOW/AIRWeb Workshop on Web Quality (WebQuality '12)*. ACM, New York, NY, 11–18.

[3] Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting Quality Flaws in User-Generated Content: The Case of Wikipedia. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 981–990.

[4] Ofer Arazy, Aron Lindberg, Mostafa Rezaei, and Michele Samorani. 2019. The Evolutionary Trajectories of Peer-Produced Artifacts: Group Composition, the Trajectories' Exploration, and the Quality of Artifacts. *MIS Quarterly* (Dec. 2019).

[5] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works and How You Can Be a Part of It.* No Starch Press, San Francisco, CA.

[6] Susan Biancani. 2014. Measuring the Quality of Edits to Wikipedia. In *Proceedings of The International Symposium on Open Collaboration (OpenSym '14)*. ACM, New York, NY, USA, 33:1–33:3.

[7] Joshua E. Blumenstock. 2008. Size Matters: Word Count as a Measure of Quality on Wikipedia. In *Proceeding of the 17th International Conference on World Wide Web - WWW '08*. ACM Press, Beijing, China, 1095.

[8] Paul-Christian Bürkner. 2017. Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software* 80, 1 (Aug. 2017), 1–28.

[9] Jaime S Cardoso, Jaime Cardoso, and Inescporto Pt. 2007. Learning to Classify Ordinal Data: The Data Replication Method. *Journal of Machine Learning Research* 8 (2007), 37.

[10] Kaylea Champion and Benjamin Mako Hill. 2021. Underproduction: An Approach for Measuring Risk in Open Source Software. *IEEE International Conference on Software Analysis, Evolution and Reengineering* (Feb. 2021). arXiv:2103.00352 [cs.SE]

[11] Hasok Chang. 2004. *Inventing Temperature.* OUP, Oxford.

[12] Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality Assessment of Wikipedia Articles Without Feature Engineering. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries (JCDL '16)*. ACM, New York, NY, USA, 27–30.

[13] Gregory Druck, Gerome Miklau, and Andrew McCallum. 2008. Learning to Predict the Quality of Contributions to Wikipedia. In *WikiAI*. 6.

[14] Andrea Forte and Amy Bruckman. 2005. Why Do People Write for Wikipedia? Incentives to Contribute to Open-Content Publishing. In *Proceedings of GROUP*. 6.

[15] C. A. E. Goodhart. 1984. Problems of Monetary Management: The UK Experience. In *Monetary Theory and Practice: The UK Experience*, C. A. E. Goodhart (Ed.). Macmillan Education UK, London, 91–121.

[16] Aaron Halfaker. 2017. Interpolating Quality Dynamics in Wikipedia and Demonstrating the Keilana Effect. In *Proceedings of the 13th International Symposium on Open Collaboration (OpenSym '17)*. Association for Computing Machinery, New York, NY, USA, 1–9.

[17] Aaron Halfaker and R Stuart Geiger. 2020. ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. 4, 148 (Oct. 2020), 37.

[18] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. 2018. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, New York.

[19] Brent Hecht and Darren Gergle. 2010. The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, Atlanta, Georgia, USA, 291–300.

[20] Dariusz Jemielniak. 2014. *Common Knowledge?: An Ethnography of Wikipedia.* Stanford University Press.

[21] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (Sept. 2016). arXiv:1609.05807 [cs, stat]

[22] Rafal Kocielnik, Os Keyes, Jonathan T. Morgan, Dario Taraborelli, David W. McDonald, and Gary Hsieh. 2018. Reciprocity and Donation: How Article Topic, Quality and Dwell Time Predict Banner Donation on Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 1–20.

[23] Florian Lemmerich, Diego Sáez-Trumper, Robert West, and Leila Zia. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. ACM, New York, NY, USA, 618–626.

[24] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* 4, 4 (Dec. 2017), 43.

[25] Richard McElreath and an O'Reilly Media Company Safari. 2018. *Statistical Rethinking.*

[26] Amanda Menking, Ingrid Erickson, and Wanda Pratt. 2019. People Who Can Take It: How Women Wikipedians Negotiate and Navigate Safety. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–14.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830.

[28] Phoebe Ayers, Charles Matthews, and Ben Yates. 2008. *How Wikipedia Works.* No Starch Press.

[29] Narun Raman, Nathaniel Sauerberg, Jonah Fisher, and Sneha Narayan. 2020. Classifying Wikipedia Article Quality With Revision History Networks. In *Proceedings of the 16th International Symposium on Open Collaboration (OpenSym 2020).* Association for Computing Machinery, New York, NY, USA, 1–7.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet Classifiers Generalize to ImageNet?. In *International Conference on Machine Learning.* PMLR, 5389–5400.

[31] Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2021. A Taxonomy of Knowledge Gaps for Wikimedia Projects (Second Draft). *arXiv:2008.12314 [cs]* (Jan. 2021). arXiv:2008.12314 [cs]

[32] Soumya Sarkar, Bhanu Prakash Reddy, Sandipan Sikdar, and Animesh Mukherjee. 2019. StRE: Self Attentive Edit Quality Prediction in Wikipedia. *arXiv:1906.04678 [cs]* (June 2019). arXiv:1906.04678 [cs]

[33] Manuel Schmidt and Eva Zangerle. 2019. Article Quality Classification on Wikipedia: Introducing Document Embeddings and Content Features. In *Proceedings of the 15th International Symposium on Open Collaboration (OpenSym '19).* Association for Computing Machinery, New York, NY, USA, 1–8.

[34] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A. Evans. 2019. The Wisdom of Polarized Crowds. *Nature Human Behaviour* 3, 4 (April 2019), 329–336.

[35] Marilyn Strathern. 1997. 'Improving Ratings': Audit in the British University System. *European Review* 5, 3 (July 1997), 305–321.

[36] Nathan TeBlunthuis, Tilman Bayer, and Olga Vasileva. 2019. Dwelling on Wikipedia: Investigating Time Spent by Global Encyclopedia Readers. In *OpenSym '19, The 15th International Symposium on Open Collaboration.* Skövde, Sweden, 14.

[37] Nathan TeBlunthuis, Benjamin Mako Hill, and Aaron Halfaker. 2021. Effects of Algorithmic Flagging on Fairness: Quasi-Experimental Evidence from Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 56:1–56:27. arXiv:2006.03121

[38] Francesca Tripodi. 2021. Ms. Categorized: Gender, Notability, and Inequality on Wikipedia. *New Media & Society* (June 2021), 14614448211023772.

[39] Maja van der Velden. 2013. Decentering Design: Wikipedia and Indigenous Knowledge. *International Journal of Human–Computer Interaction* 29, 4 (March 2013), 308–316.

[40] W. N Venables, Brian D Ripley, and W. N Venables. 2002. *Modern Applied Statistics with S.* Springer, New York.

[41] Morten Warncke-Wang, Vladislav R. Ayukaev, Brent Hecht, and Loren G. Terveen. 2015. The Success and Failure of Quality Improvement Projects in Peer Production Communities. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15).* Association for Computing Machinery, New York, NY, USA, 743–756.

[42] Morten Warncke-Wang, Dan Cosley, and John Riedl. 2013. Tell Me More: An Actionable Quality Model for Wikipedia. In *Proceedings of the 9th International Symposium on Open Collaboration (WikiSym '13).* Association for Computing Machinery, New York, NY, USA, 1–10.

[43] Ark Fangzhou Zhang, Danielle Livneh, Ceren Budak, Lionel P. Robert, and Daniel M. Romero. 2017. Crowd Development: The Interplay between Crowd Evaluation and Collaborative Dynamics in Wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (Dec. 2017), 1–21.

[44] Shiyue Zhang, Zheng Hu, Chunhong Zhang, and Ke Yu. 2018. History-Based Article Quality Assessment on Wikipedia. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp).* 1–8.

Quantifying the Gap: A Case Study of Wikidata Gender Disparities

Charles Chuankai Zhang
zhan6914@umn.edu
GroupLens Research
University of Minnesota
Minneapolis, Minnesota, USA

Loren Terveen
terveen@umn.edu
GroupLens Research
University of Minnesota
Minneapolis, Minnesota, USA

ABSTRACT

Much prior research has found *gender bias* in peer production systems like Wikipedia and OpenStreetMap. This bias affects both women's participation in these platforms and content about women on these platforms. We investigated the gender content gap in Wikidata, where less than 22% of items that represent people are about women. We asked: what is the *source* of this bias? Specifically, does it *originate from the actions of Wikidata editors* or *from external factors*; that is, does it simply reflect existing real world gender bias? We conducted a quantitative case study that found: (i) the most popular categories of people included in Wikidata represent male-dominant professions, such as American football; (ii) within a selected set of professions where we could obtain gender distribution data, Wikidata is no more biased than the real world: men and women are included at similar percentages, and the quality of items representing men and women also is similar. We provide possible explanations for our findings and implications for addressing the Wikidata content gap.

CCS CONCEPTS

• **Human-centered computing → Wikis**.

KEYWORDS

Wikidata, peer-production, structured data

ACM Reference Format:
Charles Chuankai Zhang and Loren Terveen. 2021. Quantifying the Gap: A Case Study of Wikidata Gender Disparities. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3479986.3479992

1 INTRODUCTION

Wikidata is a Wikimedia project that serves as "a free and open knowledge base that can be read and edited by both humans and machines".[1] It stores structured data about real world objects and concepts. It is widely used both by Wikimedia projects – particularly Wikipedia – and many other sites and services, such as Google, Quora and Musicbrainz. One exemplary use of Wikidata in Wikipedia is to provide data for *infoboxes*; Figure 1 shows an example infobox and its corresponding Wiki markup, which indicates data to be fetched from Wikidata.

Given Wikidata's role as a knowledge repository used by a variety of different sources, questions about the nature and quality of its information are important. Any problems or biases in Wikidata's representation of knowledge may be propagated to search engines, question answering platforms, or other online communities that use Wikidata as a source of reference or ground truth. We are particularly interested in an issue that plagues many peer production communities including (as detailed below) Wikipedia and OpenStreetMap: *gender*[2] *bias*, the under representation of content about women compared to content about men. In line with previous studies, we found that only 22% of Wikidata items that represent people are about women. The goal of our research is to begin to identify the *source of this bias*. Identifying the source of bias is important because different sources may require different remedies.

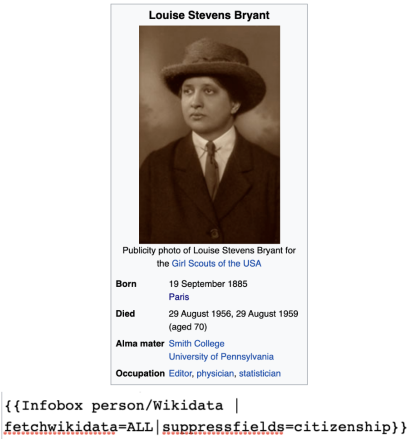
Most generally, bias can originate *from the actions of Wikidata editors* or *from external factors* (or obviously, from a combination of the two). Perhaps Wikidata editors tend to add items about men proportionally more often than items about women; or more subtly, maybe they favor adding content about categories of people where men dominate (say, popular American sports) rather than where women dominate or the genders are equally represented. On the other hand, perhaps the manifest gender bias in Wikidata content merely reflects real world biases. That is, due to discrimination and systemic biases, maybe women are underrepresented in the kinds of activities and achievements that lead people to be considered "notable" enough to be represented in Wikidata: "The entity must be notable, in the sense that it can be described using serious and publicly available references[3]." This article defines several other criteria to determine whether an entity is acceptable for representation in Wikidata. However, at least for items representing people, our reading is that notability clearly is required. Therefore, since inclusion in Wikidata is presumed to indicate that a person is "notable" (and lack of inclusion may indicate that a person is **not** "notable"), gender biases in Wikidata can lead consumers of Wikidata to form incorrect perceptions about the comparative notability of women and men.

[1] https://www.wikidata.org/wiki/Wikidata:Main_Page

[2] As we detail below, nearly all Wikidata items that represent people have a gender of either female or male, so in this paper we consider only these two genders.

[3] https://www.wikidata.org/wiki/Wikidata:Notability

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479992



Louise Stevens Bryant

Publicity photo of Louise Stevens Bryant for the Girl Scouts of the USA

Born	19 September 1885 Paris
Died	29 August 1956, 29 August 1959 (aged 70)
Alma mater	Smith College University of Pennsylvania
Occupation	Editor, physician, statistician

```
{{Infobox person/Wikidata |
fetchwikidata=ALL|suppressfields=citizenship}}
```

Figure 1: An example of Wikipedia infobox powered by Wikidata

As we prepared to investigate these possibilities, we realized that we needed to think about the representation of men and women *within specific professions* as people are likely to be recognized as *notable* for the accomplishment in the field they work on. In particular, we needed two types of data:

- The overall distribution of men and women in different professions; for this, we used United States[4] Bureau of Labor Statistics.
- The assessed "notability" of individuals in different professions; for this, we used lists of award winners for a selected set of professions.

After obtaining these datasets, we could formulate a guiding **research question** for this study:

> *To what extent is Wikidata* ***reflecting real world gender bias*** *vs.* ***introducing additional gender bias****?*

In brief, we found that:

- Wikidata editors "over sample" male-dominated professions such as American football and baseball.
- However, within a selected set of professions for which we obtained overall gender distributions and external "notability" assessments, Wikidata gender distribution is no more biased than the real world.
- Moreover, the quality of Wikidata items representing women and items representing men are equivalent.

The rest of this paper is organized as follows. First, we introduce and illustrate some key Wikidata concepts. Second, we summarize related work focusing on gender disparities in online communities. Third, we elaborate on our analytic framework and describe our data and methods. We then present our results, and conclude with a discussion of the implications of our results for future research and remedies to the observed gender gap.

2 WIKIDATA CONCEPTS AND TERMINOLOGY

Wikidata objects that represent real world entities like people are called *items*. *Label* is the most common name that an item is known by, and *description* is a short piece of information that describes the item[5]. For every item, the major bulk of its information and characteristics are stored as a list of *Statements*. A statement consist of two parts: a *claim* that the item has certain characteristics and a list of references that back up this claim. The most common form of a claim is a *property-value* pair that assigns one or multiple values to a certain property. For example, the Wikidata object[6] referred to in Figure 1 includes the following statements:

instance of **human**
sex or gender **female**
occupation **publicist**
writer
editor
physician
statistician
medical writer

The example above has three claims. The first two claims are formed by one-property-one-value pairs while the third claim is an occupation property that pairs with six values.

In this paper, we analyze only Wikidata items that are *instance of* **human**, i.e., items directly representing people. There are other types of content that could be considered to be "related" more directly to either men or women (for example, see [26] and [31]). However, since we are using real world data about gender composition of people within professions as a baseline for comparison, for our purposes it makes sense to consider only Wikidata items that represent people.

3 RELATED WORK

We report here on our research using quantitative methods to investigate the sources of gender bias in the Wikidata peer production system. Thus, our discussion of prior research focuses on work in this context that applies similar methods. However, it is worth noting that other work takes a more conceptual and theoretical approach to the issue of gender bias in online communities, including peer production systems. Some of the most relevant strands of this work [10, 32] draw on feminist HCI [3] to critique the underlying epistemological and procedural foundations of communities like Wikipedia. This work is not directly relevant to our current study, but it offers alternative perspectives aimed to create a more pluralistic and inclusive community and content, thus addressing gender bias at a fundamental level.

Prior research on peer production systems has found significant gender gaps in participation and in content coverage. Much work has focused on Wikipedia. By 2010, studies had begun to appear that found that women were a small minority of Wikipedia contributors. Glott et al. [13] conducted a survey finding that less than 13% of Wikipedia editors are women (although a revised analysis suggested that the number might be around 16% [20]). Several

[4] Analyzing only United States data is a limitation of our study. We explain why we did this in the section *Data and Methods*

[5] Since Wikidata is language independent, an item can have labels and descriptions in multiple languages.

[6] https://www.wikidata.org/wiki/Q24455644

quantitative studies found similar results and added a further finding: women were even underrepresented among the most active Wikipedia editors [1, 26]. Cabrera et al. [5] found a gender gap in participation in article talk pages, where issues concerning article content are raised and discussed. Hargittai and Shaw [18] summarized panel survey data to conclude that the most likely contributors to Wikipedia are highly skilled men. Gender gaps in participation have been attested in other populations like OpenStreetMap contributors [9, 11], StackOverflow contributors [36, 39] and open source software developers [12].

What factors lead to this large participation disparity? Prior research discovered both internal and external reasons. First, there is a significant difference between men and women in terms of online behavior. Iosub et al. [22] suggests that women Wikipedia contributors communicate in a more social and emotional manner and that women contributors are more relationship-oriented. Laniado et al. [27] found that women editors tend to communicate in a more positive tone. It is supported by the finding that distaste of high level of debate in contribution process and certain tasks like deletion are also reasons why women editors turn away from Wikipedia [4, 7]. Meanwhile, external factors are also examined in order to understand how the environment and culture of a platform contributes to the gender disparity in editors. Through an interview study, Menking and Erickson [30] found that women editors avoid certain kinds of areas or tasks that involve too much drama and stress. Organizational tensions in sociocultural norms may also cause Wikipedia women editors to experience isolation and emotional exhaustion [8]. Lir [29] analyzed the participation process and discovered pre-edit and post-edit barriers that deters women from contributing to Wikipedia.

Generally, gender gaps are a consequence of the *culture, dynamics, and values of online communities* [33]. The various types of gender gaps cause different types of harms. A *contributor gap* often leads to a *content gap* since women and men overall may differ in their interests and specializations [6, 9], and thus the types of content they create and edit. Specifically, previous research showed that Wikipedia's editor gender gap was associated with poor coverage and quality of topics that appealed more to women than men [26]. Other research found that Wikipedia biographies covered a much higher proportion of men than women, but the women who were included tended to be more notable than men, due to a hypothesized "glass ceiling" effect [38]. This research also found that articles about men and women covered different types of topics; for example information about relationship and family was more likely to be included in Wikipedia articles about women [37], while cognition related content was more likely to be included in articles about men [14]. In addition, OpenStreetMap and Google MapMaker both were shown to have gender biases in the types of places they included [34].

In previous research on gender disparity, Wikidata was used as a data source for measuring disparity in Wikipedia. For example, Klein et al. [24] built a Wikipedia gender gap indicator using Wikidata as a data source. An in-depth analysis of claim coverage and Wikidata human items by place of birth and citizenship was conducted to help them build up the indicator. In a case study on members of the European Parliament, Hollink et al. [21] compared the number of claims and family/relationship related properties between men and women Wikidata items. They found only a small difference in number of properties. They also found no evidence indicating family/relationship related properties shows up more in Wikidata items representing women, in contrast to the result from Wikipedia.

Thus far there is no systematic account of the gender gap in Wikidata; specifically, there has not been an investigation of the causes of the gap. This is important because different causes may require different solutions. If the gap originates from actions of Wikidata editors, then solutions would have to focus on the editors, for example, the composition of the editor population or tools designed to change editor behavior. On the other hand, if the gap primarily reflects existing real world biases, then solutions might require new policies to "over sample" external data to fight against these biases.

4 DATA AND METHODS

As we have explained, the Wikidata gender gap could originate *from the actions of Wikidata editors* or *from external factors*. In other words, are Wikidata contributors *causing the gap* or *reflecting an existing gap*? To answer this question, we need to compare Wikidata data to external data.

We realized we needed data organized by *professions*: as Kay et al. [23] noted: the “portrayal of occupations" is a “topic of societal importance that has recently received attention and efforts to ameliorate biases". Moreover, different professions have different gender distributions and different barriers to advancement, that is, what types of people become recognized as “notable". Therefore, organizing data by profession let us address several specific questions:

(1) *How does the Wikidata gender distribution within a profession compare to the overall gender distribution within that profession?* To answer this question, we need a dataset of gender distribution by profession.
(2) *How does the Wikidata gender distribution within a profession compare to that profession's “notability" assessments?* To answer this question, we need a dataset of people recognized as notable within various professions, along with the gender distribution of the people so recognized.
(3) *Which types of professions have most coverage in Wikidata? Are these professions more balanced or biased in gender representation?* To answer this question, we need a dataset of Wikidata items that represent people, where the person's profession and gender also are provided.

We faced several challenges in collecting the datasets that led us to take an iterative approach to defining and then *refining* the datasets. We narrate these challenges and explain the assumptions we made as we describe each dataset.

4.1 Gender Distribution by Profession: BLS Dataset

For overall gender distribution within professions, we used the United States Bureau of Labor Statistics' Current Population Survey dataset[7] as of the year 2019. For our purposes, this let us calculate the gender distribution[8] within a large number of professions.

Analyzing only United States data is a limitation. We accepted this limitation due to the availability of a high-quality data source, which is not duplicated globally. Further, this dataset has been used in previous research [23, 24] to serve as a ground truth of gender representation, again with analysis limited to the United States.

4.2 Gender Distribution by Profession: Wikidata Dataset

4.2.1 Initial Dataset Construction. We used the October 19, 2019 Wikidata data dump. We first extracted all Wikidata items that represent people, that is, items whose *instance of* (P31) property had the value *human* (Q5). This resulted in 5,477,414 items, comprising 8.5% of all items in the dump. We next filtered to include only items that had values for four properties necessary for our analysis: *sex or gender* (P21), *occupation* (P106), *date of birth* (P569) and *country of citizenship* (P27). This left us with 2,513.518 items, or just under 46% of all the human Wikidata items. We further required certain values for these properties:

- *country of citizenship* must be *United States of America* (Q30); this was necessary for comparison with the BLS dataset.
- *date of birth* (P569) had to be at least as recent as 1950; we did this for comparison with the two external datasets, as most people of this age are still employed (and thus represented in the BLS data) and have had the opportunity to become recognized as “notable" in their profession.[9]

This final filter let us with a dataset consisting of 141,562 Wikidata items representing people with US citizenship, born after 1950, with a known gender and profession.

4.2.2 Organizing by Profession. We next had to group the items in this dataset by *profession*. We initially limited ourselves to professions with more than 100 items; this yielded 133 professions. These professions and their counts are listed in the Appendix. Next, we needed to match those professions to the profession listed in the census dataset. Like others before us [23], we encountered the problem of polysemy; many BLS categories cover multiple distinct professions that are distinguished in Wikidata. For example, the BLS ’Athletes, coaches, umpires, and related workers’ profession corresponds to 42 distinct Wikidata professions, such as “American football player", “baseball player", “basketball official", and “sports commentator". Only seven of the 133 Wikidata professions with at least 100 items had a 1-to-1 match with BLS categories. Five of these seven professions were academic professions: chemistry, computer science, economics, psychology, and sociology. We selected these academic professions for our notability dataset to create a focused baseline for comparison.

However, we still had one more step for the five selected academic professions. Some Wikidata professions may be subclasses of others, represented using the *subclasses of* (P279) property. For example, a *theoretical chemist* (Q85519878) is a subclass of *physical chemist* (Q16744668) and *physical chemist* is a subclass of *chemist*. Thus, someone whose profession is *theoretical chemist* should be included among *chemists* in our dataset. Therefore, for each of the five selected professions we expanded the subclass hierarchy until we reached leaf nodes. Then for each human item in our dataset, if its *occupation* property included any profession in the class hierarchy rooted at one of the selected professions, we assigned the item to that profession.

Table 1: Academic professions and their corresponding associations

Profession	Society & Association
chemist	American Chemical Society[11]
psychologist	American Psychological Association[12]
sociologist	American Sociological Association[13]
computer scientist	Association for Computing Machinery[14]
economist	American Economic Association[15]

4.3 Notability Dataset for Five Academic Professions

Finally, since people are supposed to be notable to be represented in Wikidata, we needed to obtain external datasets of notable people within the five selected academic professions. We believe that professional society’s award recipients are the best source for this. To be clear, we are not saying anything about whether this type of recognition is fair or unbiased; we simply are saying that it reflects a profession’s assessment of the notable people within its field.

Table 1 lists for each of the five selected academic professions the professional society from which we obtained lists of award recipients. We collected this information in September 2019. We wanted to “synchronize" the notability datasets with Wikidata and BLS datasets. Recall that the BLS dataset deals with currently employed people, and we limited the Wikidata dataset to people born after 1950. We decided that by the time people were 30, they were almost certain to be employed, and they had some chance of having received recognition in their field. Therefore, we included only award recipients from the professional societies who received their award beginning in 1980.

Finally, we determined the gender of award recipients in two ways. First, if an award recipient was included in Wikidata, we retrieved their gender from Wikidata. (We report Wikidata coverage of the notability datasets below.) Otherwise, we used the gender-guesser python library[10] which has a 97.34% gender identification accuracy on Wikidata dataset [25]. The result of this tool for any given name will be one of *unknown (name not found), andy (androgynous), male, female, mostly_male, or mostly_female.* We used this tool and kept only the *male* and *female* classification results. The rest of the data were hand labeled using different sources such as Google and Wikipedia. We realize that this procedure may make incorrect gender classifications, and this is a limitation of our approach.

[7] https://www.census.gov/programs-surveys/cps.html

[8] as noted previously, we limited ourselves to men and women genders due to data availability issues.

[9] We also observed that some items do not have an exact birth date. For example, some people are listed only as born in the “*20th century*"; in this case, the data in the dump is *+2000-00-00T00:00:00Z*. We filter out these items, too.

[10] https://pypi.org/project/gender-guesser/

Table 2: List of genders and their proportion in Wikidata dataset and 5 profession dataset

Wikidata QID	Gender	Wikidata	5 Professions
Q6581072	female	24.2%	19.4%
Q6581097	male	75.6%	80.5%
Q1052281	transgender female	0.10%	0.06%
Q2449503	transgender male	0.003%	0
Q48270	non-binary	0.002%	0
Q301702	two-spirit	7e-6	0
Q1097630	intersex	3e-5	0
Q505371	agender	1e-5	0
Q189125	transgender person	7e-6	0

Table 2 shows genders listed in Wikidata and the five selected academic professions and their proportions. *Male* and *Female* genders account for more than 99.8% of the data. Therefore, we only were able to analyze distribution of these two genders. Future work is necessary to obtain sufficient data to examine biases across a wider range of genders.

5 RESULTS AND ANALYSIS

We next present our results, organized around potential Wikidata gaps (relative to external data) in *coverage* and *quality*.

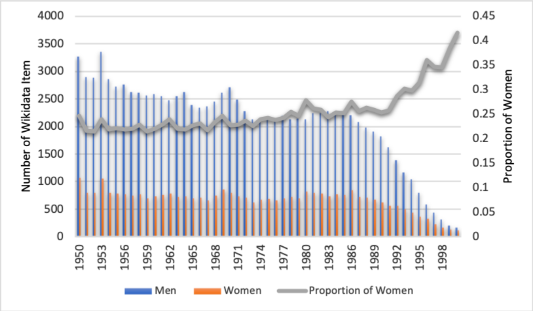


Figure 2: Number of items and proportion of women per year

5.1 Coverage Gap

Figure 2 shows the number of human items and gender proportion per year in the Wikidata dataset. Blue bars represent the number of Wikidata items about men born in each year, and the orange bars represent the number of items about women born in that year. The line on the chart shows the gender proportion trend over time. From the graph, we observed that the proportion of Wikidata items about women ranges between 0.2 and 0.25 for birth years 1950 to 1990 and has increased steadily since then, reaching 0.4178 for the 2000 birth year.

To investigate further, we list the five professions with the most items by five-year blocks. Table 3 shows the top professions every five years and their percentage in the dataset. We can observe that only eight professions appear in all the top five ranking lists. Four of them are sports related professions dominated by men: American football player (99.83%), basketball player (86.98%), baseball player (99.78%) and association football player (85.54%). As for the other professions, politician (76.85%) is also heavily biased towards men, while actor (51.12%), singer (54.41%) and writer (59.13%) are more equally represented in Wikidata. While 1392 professions occurred in the Wikidata dataset, the top five professions cover at least 30% of the data.

Thus, we can articulate an obvious gender coverage bias immediately: many of the professions most commonly represented in Wikidata are male-dominated. This in turn will skew the overall gender distribution in favor of men.

We can make several conjectures concerning the increase in representation of women among people born after 1990. First, the number of people born in this time span included in Wikidata decreases significantly. For example, several thousand people are represented in Wikidata for each 1980s birth year, but fewer than 300 for birth year 2000. This makes sense, as people who are only in their 20s have had less chance to become "notable". Second, among people born in the 1990s who are represented in Wikidata, non-sports related professions – which are much less male dominated – make up a significantly larger proportion. For example, for people born between 1986 and 1990, four of the five top professions are sports related, male-dominated, and they collectively account for nearly 46% of Wikidata human items. The one non-sports profession, *Actors*, which has virtually equal gender distributions, accounted for just under 9% of human items. However, for people born between 1996 and 2000, there are three (male-dominated) sports professions in the top 5, which collectively account for just under 29% of human items, while *Actors* is joined in the top 5 by *Singers* – another profession with close to equal gender distribution – and these two professions together accounted for over 23% of human items.

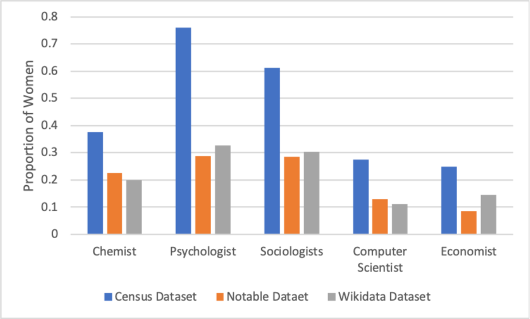


Figure 3: Proportion of women in 5 academic professions across three datasets

We next compared Wikidata gender distribution within the five selected academic professions to the gender distribution in the profession as a whole (BLS data) and in professional societies' notability assessments. Figure 3 shows the proportion of women in the five academic professions in each of our three datasets. We first

[11] https://www.acs.org/content/acs/en/awards.html
[12] https://www.apa.org/about/awards
[13] https://www.asanet.org/about/awards
[14] https://awards.acm.org/
[15] https://www.aeaweb.org/about-aea/honors-awards

Table 3: Top 5 professions with most data in every five years

Year	Top 1	Top 2	Top 3	Top 4	Top 5	Proportion
1951 - 1955	politician	actor	American football player	writer	baseball player	30.5%
1956 - 1960	politician	American football player	actor	baseball player	writer	32.0%
1961 - 1965	American football player	actor	politician	baseball player	basketball player	34.6%
1966 - 1970	actor	American football player	politician	baseball player	basketball player	32.9%
1971 - 1975	American football player	actor	baseball player	politician	basketball player	35.9%
1976 - 1980	American football player	actor	basketball player	baseball player	singer	40.7%
1981 - 1985	American football player	actor	basketball player	baseball player	association football player	47.3%
1986 - 1990	American football player	basketball player	actor	association football player	baseball player	54.8%
1991 - 1995	American football player	basketball player	association football player	actor	baseball player	57.1%
1996 - 2000	actor	association football player	basketball player	American football player	singer	52.2%

observed that both Wikidata and professional societies' notability assessments include a much higher proportion of men than in the profession as a whole. On the other hand, the graph suggests that Wikidata gender distributions are no more biased than the professional societies. There is a high within-pair correlation (r = 0.923) between the gender proportion of notable dataset and Wikidata dataset so we have sufficient statistical power to run a paired t-test on this small sample size (n=5). The t-test result between the Wikidata dataset and the notable dataset shows that there is not a significant difference between them (p = 0.414).

The previous analysis let us compare the overall gender distributions of the Wikidata and the notability datasets. We also examined the *specific* coverage of the notability datasets – that is, the people recognized by the five professional societies – in Wikidata. For every person in the notability datasets, we checked whether they were represented in Wikidata. We used their name as a query, and considered ourselves to have found a match if and only if exactly one human item with a matching profession is found. For example, the 1980 Association of Computing Machinery (ACM) Turing Award winner was Tony Hoare. The ACM lists his name as C. ANTONY ("TONY") R. HOARE[16]. Using this as a query to Wikidata returns exactly one item – *Tony Hoare (Q92602)* – and this item is an *instance of* a *human*, and its *occupation* property includes the value *Computer Scientist*. Therefore, this is a successful match. The 1987 Turing Award Winner was John Cocke[17]. The ACM lists his name as "John Cocke". Using this as a query to Wikidata returns a large number of Wikidata items, but only one – *John Cocke (Q92632)* – is an *instance of* a *human*, and has an *occupation* property that includes the value *Computer Scientist*. Therefore, this too is a successful match. On the other hand, many recipients of some other ACM awards, such as Distinguished Member, are not found in Wikidata at all. A final note is that we once again use the Wikidata profession hierarchy in this process. So for example, if someone honored by the American Chemical Society was listed as a *theoretical chemist* in Wikidata, we treat this as a match, too.

Figure 4 shows the results of this analysis. It shows that there is a large difference in coverage among professions. More than 90% of economists recognized by the American Economic Association are represented in Wikidata, while only about 40% of chemists recognized by the American Chemical Society are included. In no profession is there a significant difference between the proportion of men and women award recipients who are represented in Wikidata. In absolute terms, in four of the five professions, a higher proportion of women award recipients is included in Wikidata.

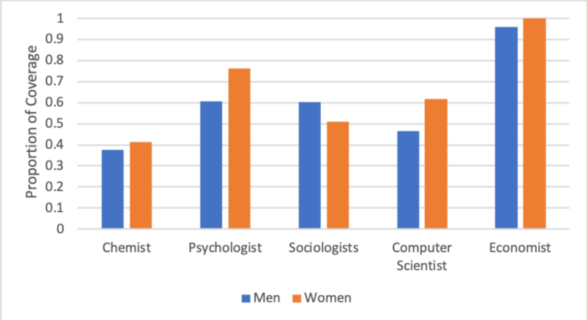


Figure 4: Notable dataset Coverage in Wikidata

5.2 Quality Gap

In this section, we investigate whether there is a difference in the *quality* of Wikidata items that represent men and items that represent women. We limit our analysis to human items within the five academic professions since items within these professions are likely to be characterized by a similar set of properties. In other words, it is easier to compare the quality of two academic professionals than to compare an academic professional to an actor or athlete. In our analysis, we use both specific quality metrics and other factors that are associated with quality.

5.2.1 Direct Metrics. We use the Objective Revision Evaluation Service (ORES) [17] to evaluate the quality of Wikidata items. ORES is a service provided by the Wikimedia Foundation that predicts edit quality and assists content moderation for various Wikimedia projects [16].

A Wikidata data dump includes the most recent revision of each Wikidata item at the time the dump was created. We extracted revision IDs from the Wikidata dataset and used ORES' quality evaluation API to estimate item quality. ORES uses the Wikidata quality assessments, which range from A (highest) to E (lowest)[18]. Specifically, ORES returns the probability that an item should be classified at each level. We then used the weighted sum formula (Formula 1) proposed by Halfaker [15] to compute a single score ranging from highest quality (4) to lowest (0). An item would be

[16] https://amturing.acm.org/award_winners/hoare_4622167.cfm
[17] https://amturing.acm.org/award_winners/cocke_2083115.cfm
[18] https://www.wikidata.org/wiki/Wikidata:Item_quality

scored a 4 if ORES predicted a 100% probability that the item should be classified as quality level A, and would be scored a 0 if ORES predicted a 100% probability that the item should be classified as quality level E. Figure 5 shows an an example of a prediction given by ORES to a particular item; using the weighted sum formula, the item would receive a score of 2.9901, which corresponds to quality level B.

$$\begin{aligned}\text{Weighted Sum} = {} & 4 \times \text{P(item is of quality A)}+ \\ & 3 \times \text{P(item is of quality B)}+ \\ & 2 \times \text{P(item is of quality C)}+ \\ & 1 \times \text{P(item is of quality D)}+ \\ & 0 \times \text{P(item is of quality E)}\end{aligned} \tag{1}$$

```
"score": {
  "prediction": "B",
  "probability": {
    "A": 0.07928625288897875,
    "B": 0.8437658868814822,
    "C": 0.06826526593440399,
    "D": 0.005158008654169078,
    "E": 0.0035245856409659073
  }
}
```

Figure 5: An example of ORES quality score prediction

As shown in the leftmost part of Figure 6 (and confirmed by the t-test result in Table 4), there is no significant difference in the ORES scores between Wikidata items representing women and items representing men. However, we want to take a closer look at a few specific important features, to see if any of them exhibited significant differences between men and women items.

- The number of *claims* constitutes the total amount of information about a Wikidata item.
- *Labels* and *descriptions* are multilingual, so the more of each, the better the representation of an item in multiple languages.
- *Sitelinks* link to other Wikimedia projects, so more sitelinks means the item is better connected to the larger Wikimedia ecosystem.

The remainder of Figure 6 shows box plots four ORES scores. We performed an independent t-test on these features. Table 4 shows the median and mean of the features and the resulting p-values of the t-tests. Only one feature shows a statistically significant difference: Wikidata items about women have a mean of 18.35 claims, while items about men have a mean of 19.74 claims (p = 0.01).

5.2.2 Associated Factors. Previous research in Wikipedia found that editors' attention and effort correlated strongly with the quality of Wikipedia articles. For example, the numbers of revisions to an article and the number of unique authors are strong predictors of article quality [28, 35]. This makes intuitive sense: more revisions indicates more effort, while more unique editors indicates more diverse perspectives.

Therefore, we want to see whether this relationship held in Wikidata: do more revisions and more unique editors for an item correlate with higher quality scores, as computed by ORES? The answer is yes. Using Spearman correlation, we found a strong positive association strong positive correlation between the number of revisions and ORES scores (rs(1602) = .78, p < .001), and between the number of distinct editors and ORES score (rs(1602) = .73, p < .001).

Finally, we check to see whether men and women Wikidata items differed in number of revisions and number of unique editors. Independent t-tests show that there is no significant difference in either case. For number of revisions, women items have a mean of 81.6 revisions (SD=54.5) and men items have a mean of 82.6 items (SD=52.0), ns (p=0.762). For number of unique editors, women items have a mean of 38.1 (SD = 21.9) and men items have a mean of 37.3 (SD=19.), ns (p=0.523).

6 DISCUSSION

Our guiding research question is: to what extent is Wikidata **reflecting real world gender bias** vs. **introducing additional gender bias**? Our analysis suggests answers.

We found that Wikidata editors are likely to over sample male-dominated professions such as American football and baseball, thus contributing to the general predominance of items representing men over items representing women. Our analysis that focused on a set of academic professions show that the gender distribution of Wikidata is no more biased than real world notability judgments in either coverage or quality. We next discuss some possible explanations for our results, and how the structured nature of Wikidata may lead to reduced bias. We also discuss some low quality Wikidata items we observed during our data collection process, which lets us discuss the role and importance of human effort. Finally, we mention the possible role of self-focus bias and identify directions for future work.

6.1 Wikidata's Factual Basis May Reduce Bias

One notable finding of our case study is that Wikidata's *coverage* of women vs. men is no more biased than real world notability assessments within a set of academic professions. The percentage of Wikidata items representing women in these professions is comparable to the percentage of women who received awards from the corresponding professional societies. More promising is the fact that the quality of items representing men and women is equivalent. This contrasts with studies of Wikipedia, which have shown biases in content about and relevant to women [26, 31, 37].

Several factors may explain why the quality of Wikidata items about men and women is comparable. First, Wikidata data for a person consists of facts about that person, such as name, date of birth, place of birth, country of citizenship, occupation, etc. More specifically, within a profession, additional properties might be prominent. For example, for politicians, these include their political party and elected position(s) held. Providing this sort of factual information about a person is more straightforward than editing a Wikipedia article. We conjecture that it does not offer as much opportunity for gender bias – even implicit types of bias – to creep

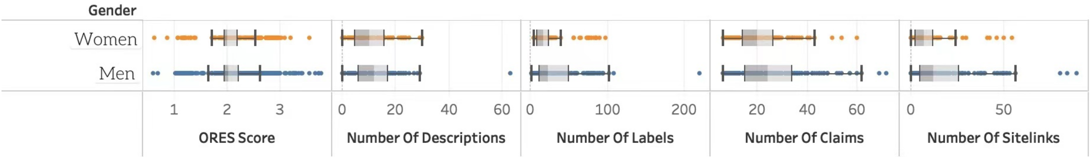


Figure 6: Box plots for ORES score and 4 features

Table 4: Summary of basic properties in men and women Wikidata pages. Statistically significant result is bolded.

	Women Wikidata Pages median / (mean)	Men Wikidata Pages median / (mean)	p-value
Number of Labels	11 (15.87)	11 (15.73)	0.887
Number of Descriptions	7 (8.78)	7 (9.07)	0.477
Number of Claims	**17 (18.35)**	**18 (19.74)**	**0.010***
Number of Sitelinks	2 (4.39)	2 (4.36)	0.962
ORES score	2.00 (2.11)	2.01 (2.13)	0.418

in, as has been shown in multiple comparisons of language used in Wikipedia biographies [2, 14, 38].

Further, much Wikidata content is added via automated bots, which import information from external data sources such as the Encyclopedia Brittanica. This also contributes to bringing in equivalent types and amounts of factual data about both women and men.

6.2 Human and bots both play vital roles

While bots are useful in bringing content into Wikidata, some exploratory analysis we did emphasizes the necessary roles of humans as well as bots. One of our filters for a *human* item to be included in our analysis is that it must have the properties gender, date of birth, country of citizenship and occupation. Without this information, it can be hard even to know which actual person an item refers to, and it obviously precludes many types of analysis. But 54% of human items did not pass this filter in our initial data collecting phase.

We encountered this problem when trying to determine whether people in our notability datasets are included in Wikidata. We sometimes found items with a matching name and perhaps a general profession such as *researcher* or *scientist* and one or two other properties, but we were not able to tell if this was the person in question. A bot might be able to import information about a person from a database, but a human Wikidata editor might be able to locate that person's website and to find and add additional important information. Future work could further explore the complementary role of human editors and bots and identify opportunities for tools to effectively combine human and automated effort.

6.3 Topical coverage and self-focus bias

A major source for the predominance of men items in Wikidata is the differential coverage of professions. Notably, three or four male-dominated sports professions are among the top five professions during each five-year interval. While it certainly is the case that some professions simply receive more attention, which makes them more likely to be covered in Wikidata, another reason may be playing a role: *self-focus bias*. Previous work has shown that contributors to peer production communities naturally enter and edit information on topics of interest to themselves [19]. Thus, *if* Wikidata editors consist mostly of men, *then* self-focus bias likely is contributing to this particular gender coverage gap. Future work to investigate the demographics of Wikidata editors would be helpful.

7 SUMMARY

We conducted a case study of the gender content gap in Wikidata. We began by noting that only 22% of Wikidata items representing people are about women. This led us to ask: was this due to existing real-world biases, or was it due to decisions of Wikidata editors? We answered this question by comparing Wikidata data to two external datasets, US Bureau of Labor Statistics data that showed the overall gender distribution within professions, and lists of award winners by a set of professional societies, which indicate who is considered “notable" within those professions. We found that Wikidata's representation of women within a set of professions was comparable to the professional societies' notability assessments, and both contained lower proportions of women than in the profession as a whole. We also observed that many of the professions with most items in Wikidata are male-dominated sports professions. Finally, we found that the quality of Wikidata items representing women was comparable to the quality of items representing men. We discussed several implications and possible next steps based on our findings.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their comments and suggestions that helped us strengthen our paper. This work was supported by the National Science Foundation(NSF) under Award No. IIS-1816348.

REFERENCES

[1] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender differences in Wikipedia editing. In *Proceedings of the 7th international symposium on wikis*

and open collaboration. 11–14.

[2] David Bamman and Noah A Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics* 2 (2014), 363–376.

[3] Shaowen Bardzell. 2010. Feminist HCI: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1301–1310.

[4] Julia B Bear and Benjamin Collier. 2016. Where are the women in Wikipedia? Understanding the different psychological experiences of men and women in Wikipedia. *Sex Roles* 74, 5-6 (2016), 254–265.

[5] Benjamin Cabrera, Björn Ross, Marielle Dado, and Maritta Heisel. 2018. The Gender Gap in Wikipedia Talk Pages. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.

[6] Shuo Chang, Vikas Kumar, Eric Gilbert, and Loren G Terveen. 2014. Specialization, homophily, and gender in a social curation site: Findings from Pinterest. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 674–686.

[7] Benjamin Collier and Julia Bear. 2012. Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 383–392.

[8] Danielle J Corple. 2016. Beyond the Gender Gap: Understanding Women's Participation in Wikipedia. (2016).

[9] Maitraye Das, Brent Hecht, and Darren Gergle. 2019. The gendered geography of contributions to OpenStreetMap: Complexities in self-focus bias. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[10] Casey Fiesler, Shannon Morrison, and Amy S Bruckman. 2016. An archive of their own: a case study of feminist HCI and values in design. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2574–2585.

[11] Z Gardner, Peter Mooney, S De Sabbata, and L Dowthwaite. 2020. Quantifying gendered participation in OpenStreetMap: responding to theories of female (under) representation in crowdsourced mapping. *GeoJournal* 85, 6 (2020), 1603–1620.

[12] Rishab A Ghosh, Ruediger Glott, Bernhard Krieger, and Gregorio Robles. 2002. Free/libre and open source software: Survey and study.

[13] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. 2010. Wikipedia survey–overview of results. *United Nations University: Collaborative Creativity Group* 8 (2010), 1158–1178.

[14] Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. 2015. First women, second sex: Gender bias in Wikipedia. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. 165–174.

[15] Aaron Halfaker. 2017. Interpolating quality dynamics in Wikipedia and demonstrating the Keilana effect. In *Proceedings of the 13th International Symposium on Open Collaboration*. 1–9.

[16] Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–37.

[17] Aaron Halfaker, Jonathan Morgan, Amir Sarabadani, and Adam Wight. 2016. ORES: Facilitating re-mediation of Wikipedia's socio-technical problems. *Working Paper, Wikimedia Research* (2016).

[18] Eszter Hargittai and Aaron Shaw. 2015. Mind the skills gap: the role of Internet know-how and gender in differentiated contributions to Wikipedia. *Information, communication & society* 18, 4 (2015), 424–442.

[19] Brent Hecht and Darren Gergle. 2009. Measuring self-focus bias in community-maintained knowledge repositories. In *Proceedings of the fourth international conference on communities and technologies*. 11–20.

[20] Benjamin Mako Hill and Aaron Shaw. 2013. The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation. *PloS one* 8, 6 (2013), e65782.

[21] Laura Hollink, Astrid Van Aggelen, and Jacco Van Ossenbruggen. 2018. Using the web of data to study gender differences in online knowledge sources: the case of the European parliament. In *Proceedings of the 10th ACM Conference on Web Science*. 381–385.

[22] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. 2014. Emotions under discussion: Gender, status and communication in online collaboration. *PloS one* 9, 8 (2014), e104880.

[23] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3819–3828.

[24] Maximilian Klein, Harsh Gupta, Vivek Rai, Piotr Konieczny, and Haiyi Zhu. 2016. Monitoring the gender gap with wikidata human gender indicators. In *Proceedings of the 12th International Symposium on Open Collaboration*. 1–9.

[25] Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*. 108–113.

[26] Shyong (Tony) K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. WP: clubhouse? An exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*. 1–10.

[27] David Laniado, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. 2012. Emotions and dialogue in a peer-production community: the case of Wikipedia. In *proceedings of the eighth annual international symposium on wikis and open collaboration*. 1–10.

[28] Andrew Lih. 2004. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. *Nature* 3, 1 (2004), 1–31.

[29] Shlomit Aharoni Lir. 2019. Strangers in a seemingly open-to-all website: the gender bias in Wikipedia. *Equality, Diversity and Inclusion: An International Journal* (2019).

[30] Amanda Menking and Ingrid Erickson. 2015. The heart work of Wikipedia: Gendered, emotional labor in the world's largest online encyclopedia. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 207–210.

[31] Amanda Menking, David W McDonald, and Mark Zachry. 2017. Who Wants to Read This? A Method for Measuring Topical Representativeness in User Generated Content Systems. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2068–2081.

[32] Amanda Menking and Jon Rosenberg. 2021. WP: NOT, WP: NPOV, and Other Stories Wikipedia Tells Us: A Feminist Critique of Wikipedia's Epistemology. *Science, Technology, & Human Values* 46, 3 (2021), 455–479.

[33] Joseph Reagle. 2013. "Free as in sexist?" Free culture and the gender gap. *first monday* (2013).

[34] Monica Stephens. 2013. Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal* 78, 6 (2013), 981–996.

[35] Besiki Stvilia, Michael B Twidale, Linda C Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-Based Encyclopedia. *ICIQ* 5, 2005 (2005), 442–454.

[36] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2012. Gender, representation and online participation: A quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*. IEEE, 332–338.

[37] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.

[38] Claudia Wagner, Eduardo Graells-Garrido, David Garcia, and Filippo Menczer. 2016. Women through the glass ceiling: gender asymmetries in Wikipedia. *EPJ Data Science* 5 (2016), 1–24.

[39] Elijah Zolduoarrati and Sherlock A Licorish. 2021. On the Value of Encouraging Gender Tolerance and Inclusiveness in Software Engineering Communities. *Information and Software Technology* (2021), 106667.

A APPENDIX: 133 WIKIDATA PROFESSIONS WITH MORE THAN 100 USA DATA ITEMS

Wikidata Qid	Occupation	Number of items
Q19204627	American football player	16090
Q33999	actor	12706
Q82955	politician	8083
Q3665646	basketball player	8019
Q10871364	baseball player	7413
Q937857	association football player	4386
Q177220	singer	4148
Q36180	writer	3391
Q639669	musician	3258
Q1930187	journalist	2762
Q40348	lawyer	2603
Q11774891	ice hockey player	2352
Q2526255	film director	1979
Q36834	composer	1939
Q488205	singer-songwriter	1830
Q488111	pornographic actor	1815
Q483501	artist	1679
Q6625963	novelist	1658
Q28389	screen writer	1516
Q11513337	athletics competitor	1412
Q11338576	boxer	1120
Q43845	businessperson	1120
Q10798782	television actor	1053
Q2252262	rapper	1024
Q11303721	golfer	993
Q13382576	rower	957
Q33231	photographer	921
Q13474373	professional wrestler	919
Q1028181	painter	909
Q11607585	mixed martial artist	877
Q4610556	model	839
Q10833314	tennis player	813
Q2309784	sport cyclist	811
Q753110	songwriter	800
Q2066131	athlete	787
Q49757	poet	774
Q15981151	jazz musician	725
Q10800557	film actor	724
Q3282637	film producer	723
Q15117302	volleyball player	719
Q5137571	basketball coach	687
Q10843402	swimmer	667
Q81096	engineer	656
Q201788	historian	614
Q183945	record producer	608
Q855091	guitarist	603
Q2405480	voice actor	578
Q19595175	amateur wrestler	513
Q2722764	radio personality	512
Q386854	drummer	501
Q189290	military officer	474
Q378622	racing driver	474
Q131524	entrepreneur	454

Continued on next page

Table 5 – *Continued from previous page*

Wikidata Qid	Occupation	Number of items
Q82594	computer scientist	453
Q170790	mathematician	413
Q482980	author	409
Q13219587	figure skater	409
Q188094	economist	406
Q3246315	head coach	401
Q212980	psychologist	341
Q715301	comics artist	335
Q130857	disc jockey	324
Q19841381	Canadian football player	312
Q2259451	stage actor	302
Q3400985	academic	298
Q245068	comedian	280
Q1238570	political scientist	273
Q41583	coach	271
Q486748	pianist	264
Q169470	physicist	253
Q4009406	sprinter	246
Q4964182	philosopher	236
Q4773904	anthropologist	235
Q4144610	alpine skier	224
Q158852	conductor	220
Q10349745	racing automobile driver	217
Q947873	television presenter	216
Q2986228	sports commentator	213
Q13381572	artistic gymnast	212
Q1114448	cartoonist	207
Q17682262	lacrosse player	207
Q15295720	poker player	206
Q1281618	sculptor	206
Q11063	astronomer	195
Q47064	military personnel	194
Q37226	teacher	188
Q193391	diplomat	187
Q1622272	university teacher	180
Q578109	television producer	180
Q2961975	business executive	178
Q593644	chemist	168
Q864503	biologist	163
Q214917	playwright	156
Q42973	architect	154
Q584301	bassist	153
Q250867	Catholic priest	152
Q16533	judge	150
Q13381863	fencer	144
Q10873124	chess player	144
Q18581305	beauty pageant contestant	144
Q15709642	snowboarder	143
Q17502714	skateboarder	139
Q3014296	motorcycle racer	136
Q222344	cinematographer	136
Q14089670	rugby union player	135
Q39631	physician	134
Q2490358	choreographer	133

Continued on next page

Table 5 – *Continued from previous page*

Wikidata Qid	Occupation	Number of items
Q6665249	judoka	132
Q644687	illustrator	129
Q16029547	biathlete	127
Q846750	jockey	126
Q5716684	dancer	125
Q3501317	fashion designer	123
Q484876	chief executive officer	119
Q17524364	water polo player	117
Q901	scientist	117
Q13561328	surfer	115
Q11631	astronaut	113
Q17125263	YouTuber	112
Q13388586	softball player	110
Q2374149	botanist	109
Q10866633	speed skater	108
Q18617021	freestyle skier	108
Q2306091	sociologist	108
Q3499072	chef	108
Q15982858	motivational speaker	107
Q2059704	television director	106
Q484188	serial killer	104
Q15253558	activist	103
Q14467526	linguist	102
Q10843263	field hockey player	102
Q13141064	badminton player	100

Implicit Visual Attention Feedback System for Wikipedia Users

Neeru Dubey
neerudubey@iitrpr.ac.in
Indian Institute of Technology Ropar
Rupnagar, Punjab, India

Amit Arjun Verma
2016csz0003@iitrpr.ac.in
Indian Institute of Technology Ropar
Rupnagar, Punjab, India

S.R.S. Iyengar
sudarshan@iitrpr.ac.in
Indian Institute of Technology Ropar
Rupnagar, Punjab, India

Simran Setia
2017csz0001@iitrpr.ac.in
Indian Institute of Technology Ropar
Rupnagar, Punjab, India

ABSTRACT

The complex collaborative structure of Wikipedia has attracted researchers from various domains, such as social networks, human-computer interaction, and collective intelligence. Yet, a few focus on the readers' perception of Wikipedia. Readers make up the majority of Wikipedia users (editors/readers), and being on the consumption side, readers play a crucial role in its sustenance. The attention patterns of users while reading an article can reveal users' interest distribution as well as content quality of the article. In this paper, we present an Attention Feedback (AF) approach for Wikipedia readers. The fundamental idea of the proposed approach comprises the implicit capture of gaze-based feedback of Wikipedia readers using a commodity gaze tracker. The developed AF mechanism aims at overcoming the main limitation of the currently used "pageview" and "survey" based feedback approaches, i.e., data inaccuracy. Moreover, the incorporation of a single-camera image processing-based gaze tracker makes the overall system cost-efficient and portable. The proposed approach can be extended to enable the research community to analyze various online portals as well as offline documents from the readers' perspective.

CCS CONCEPTS

• **Human-centered computing → Wikis**; **Collaborative and social computing**; **Collaborative and social computing systems and tools**;

KEYWORDS

Wikipedia, gaze tracking, Collaborative analysis tool, gaze dataset, readers

ACM Reference Format:
Neeru Dubey, Amit Arjun Verma, S.R.S. Iyengar, and Simran Setia. 2021. Implicit Visual Attention Feedback System for Wikipedia Users . In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3479986.3479993

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479993

1 INTRODUCTION

Wikipedia is a collaboratively developed online encyclopedia. It has been consistently ranked in the top fifteen most-visited websites as per Alexa ranking[1]. Based on the actions performed, Wikipedia users can be broadly divided into producers (editors, moderators) and consumers (readers). Over the last years, research has extensively investigated the group of producers while another group of Wikipedia users, the readers, their preferences, and their behavior have not been much studied.

Online participation research has often characterized Internet readers as non-contributors, who benefit from others' contributions but who contribute little themselves [31]. In general, non-contributing readers constitute 90% or more of any given discussion forum, online community, or social website [25]; a 2011 Readership Survey showed that the number for Wikipedia was 94%. Although often called a "participation inequality", such statistics indicate that reading constitutes the norm, whereas contribution is the anomaly (albeit a necessary one). A few pieces of research highlight the importance of readers in the Wikipedia community. The study by Antin et al. [6] claims that reading can be seen as a form of participation and is, therefore, valuable. Reading a Wikipedia article can be considered legitimate peripheral involvement through which individuals gain knowledge and move towards more active participation. The fact that a user reads an article and not edit, is a good indication of its quality, such as its reliability [3]. Lehmann et al. [23] also emphasize the importance of reading behavior analyses to characterize users' reading preferences for portal content development.

The primary reason for the lack of research focused on readers is the limited availability of suitable techniques to capture readers' attention feedback. The current popular techniques are pageviews and Surveys. Pageviews depict the total number of visits to an article during a given period, but it has several disadvantages. Firstly, the statistics do not quantify the number of unique readers, and secondly, it does not consider the time duration a reader has spent reading the article. Another widely used method is surveying the readers. Nevertheless, it also suffers from drawbacks in terms of survey process scalability and accuracy of information provided by participants. The Wikimedia Foundation has also emphasized that pageviews and surveys are not efficient measures to identify readers' perception of an article [49]. One more technique was presented by Barifah et al. [8]. They used log files as a means to

[1] https://www.alexa.com/topsites

uncover information about users and their behavior when searching for information. This study excluded fine-grained analysis due to lack of details of users' navigation patterns.

This paper proposes a novel framework to capture the Attention Feedback (AF) of Wikipedia readers. The proposed method records the implicit gaze pattern of a reader and extracts the visual feedback features. In past research, it has been well established that our eye movement is closely related to human cognition [4]. There is evidence in reading psychology literature that eye movement patterns during reading are indeed related to the textual features of the document [41]. Ajanki et al. [5] claim that when a system does not have any prior information regarding what the user wants to search, the eye movements help significantly in the search. It is the case in a proactive search, for instance, where the system monitors the reading behavior of the reader in a new document. Xu et al. [55] talk about the relevance between the human text reading pattern and their current cognitive process. The relation between eye gaze pattern and human interest has been vastly utilized in various fields [20].

The previous studies performed to capture users' gaze for online portals utilize dedicated eye trackers like Tobii [33]. The dedicated eye trackers are pretty expensive (hundreds of dollars) and mostly invasive in nature[2]. Additionally, these trackers are rarely available at users' sites. We propose to use a lightweight desk-mounted/laptop camera along with a publicly available image processing based eye tracker to capture readers' attention patterns. We embed the eye tracker and the analysis methods in a web application called "WikiRead". Usage of commodity eye-tracking solution and web application makes it user-friendly and portable. It also enables the collection of a large dataset of Wikipedia readers' AF. This dataset can open doors for the analysis of Wikipedia from a novel perspective.

The rest of the paper is structured as follows: Section 2 describes the necessary background and related works to the proposed method. Section 3 provides an overview of the proposed AF framework. In Section 4, we describe the procedure to perform attention pattern analysis. In Section 5, we discuss the system setup and the experiments performed to evaluate the proposed technique from various perspectives. In Section 6, we provide a discussion on the limitations and possible future directions. Section 7 concludes the work.

2 BACKGROUND & RELATED WORK

Eye gaze tracking is the process of estimating the direction of the sight line and the point of regard. There have been several advances in eye-tracking technology. Based on physical contact requirement with the eye tracker, the research in this field can be classified into two types; invasive eye trackers and non-invasive eye trackers. Invasive eye trackers require physical contact with the users. The most prevalent invasive trackers are Tobii, EyeLink, SMI2, Mirametrix4, and EyeTech3. They provide higher accuracy as compared to non-invasive eye trackers, but they are pretty expensive. The non-invasive eye trackers mainly work using computer vision techniques and some lightweight camera(s). Therefore, they do not require any physical contact with the user. Some recent researches include Searchgazer, xLabs Gaze Tracking, GazePointer, Ogama, OpenEyes, PyGaze, OpenGazer, TurkerGaze. The details of these non-invasive eye trackers can be referred to in [56]. We use GazeFlowAPI[3] as a gaze tracking solution due to it's efficiency and suitability for the current task.

Our eyes show a characteristic behavior composed of a series of fixations and saccades [13]. Fixation is a duration for which the eye is steadily gazing at one point or a collection of proximal points. On average, it is of a time interval of 200-250 ms. A saccade is a rapid eye movement from one fixation to the next one. The time-lined distribution of saccades and fixations reflects significant characteristics of the user's reading behavior and the object being gazed. Henny et al. [4] mentioned that eye gaze is a prominent nonverbal signal compared to pointing, body posture, and other behaviors.

We draw the motivation to use eye gaze as the medium to capture users' attention feedback, from the massive backing of literature on the relationship between eye gaze pattern and human interest [20]. Employment of eye movement data in an educational context has offered insight into how to model gaze to extract human attention patterns. The most noteworthy are the eye movement modeling examples (EMMEs), in which Jarodzka et al. used visual feedback to specifically influence gaze actions in order to enhance subjects' perception performance of medical records [17] and a biological classification task [18, 19]. Eye movement data of experts was visualized in [17] by blurring areas they did not look at, i.e., non-relevant details. Experts' gaze was visualized as yellow circles on a stimulus picture for [18]. The model example in both studies included gaze data post hoc. Orlov and Bednarik's open-source program ScreenMasker [34] created a flexible framework that visualizes gaze actions. Their gaze-contingent system overlays the on-screen stimulus with a pattern mask. It subtracts the pattern or unmasks, where the subject is looking in real-time, using gaze coordinates from the eye tracker. An NVIDIA graphics card with the CUDA architecture was used for this system [34]. Later, [35] proposed a platform integrated with a publicly available eye gaze analysis tool. The multiple plugins integrated offered an experimental center and a real-time gaze feedback option. Their system was tested and capable of running on a standard computer, but the visual search tasks were performed on images.

In this work, we design a real-time gaze-based attention feedback framework that identifies salience areas in mixed media Wikipedia articles (text, images). Wikipedia is a crowd-sourced and openly investigable source of information. As a result, ever-growing research interest has been shown to investigate various aspects of Wikipedia, and different metrics have been proposed for investigation purposes. The metrics include straightforward strategies like word count [10] to the models trained using deep learning [12] to evaluate various aspects of Wikipedia articles. Some other commonly used techniques to analyze Wikipedia articles are reference evaluation [24], editors' contribution [54], talk page evaluation [21], etc. In this work, we propose to use implicit AF from readers of

[2] Some dedicated gaze trackers like EyeLink 1000 provide non-invasive solutions but they also require high end camera and GPU processors. These requirements are rarely available at users' end

[3] https://github.com/szydej/GazeFlowAPI

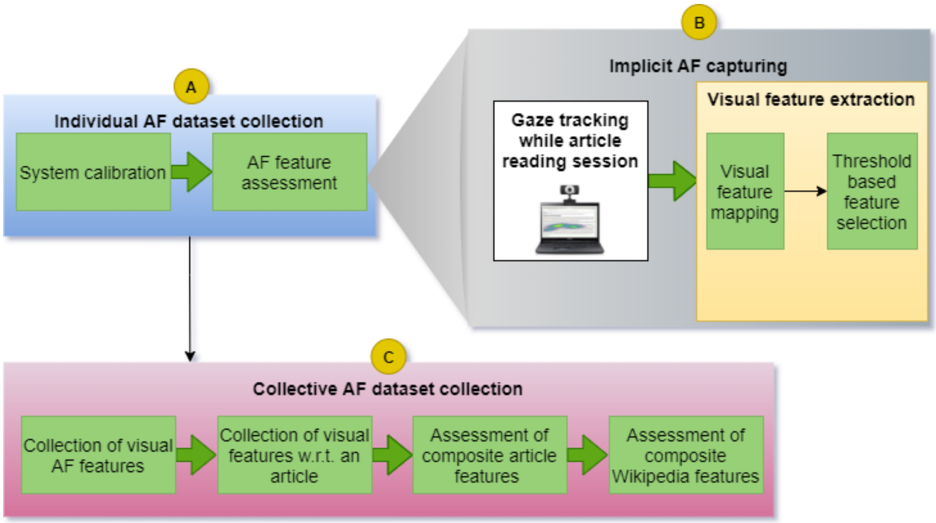

B
Implicit AF capturing
Visual feature extraction
Gaze tracking while article reading session
Visual feature mapping
Threshold based feature selection
A
Individual AF dataset collection
System calibration
AF feature assessment
C
Collective AF dataset collection
Collection of visual AF features
Collection of visual features w.r.t. an article
Assessment of composite article features
Assessment of composite Wikipedia features

Figure 1: Overall system workflow to capture readers' AF. In block A, dataset collection happens corresponding to each reading session. Block B showcases the procedure to extract visual features. In block C, readers' attention feedbacks are collected to derive composite article/Wikipedia level features.

an article to analyze that article, using the concepts of collective intelligence.

There have been few attempts in the past where readers' input has been used to analyze crowd-sourced web portals. We focus on the crowd-sourced portals because we aim to utilize the knowledge and participation of crowd (users/editors) for portal analysis. In a Wikipedia-based research [40], readers' navigation behavior was analyzed to quantify the importance of references in Wikipedia. Stackoverflow is a successful crowd-sourced QnA portal. There have been some studies on Stackoverflow, which require readers' involvement. In one such study, Peterson et al. [39] proposed a gaze-based exploratory study on readers' information-seeking behavior, esp. developers on Stackoverflow. Gaze-based input from readers has also been utilized to identify the importance of citations in Stackoverflow posts [40]. Linden et al. [50] performed a detailed study to investigate readers' behavior in selecting answers in a thread (QnA post). GitHub is another web portal that supports the collaborative effort. It is a widely-used software development platform that provides features of version control and project hosting. There have been attempts on Github also, which involves input from users. In [16], authors predicted student learning outcomes from a survey of students and teachers while using GitHub in the classroom. In [51], a survey was conducted on GitHub to get a sense of how participants assess and value team composition and diversity.

The Article Feedback Tool (AFT)[4] was a Wikimedia survey for article feedback to engage readers in the assessment of article quality. It suffered from a lack of participation and noisy data. Similarly, the other commonly used approach of pageviews [11] also suffers from data insufficiency since pageviews only provide a count of visits for an article. To the best of our knowledge, the current work presents a novel approach to capture readers' implicit feedback for Wikipedia articles using a low-cost and effective gaze-based solution.

3 SYSTEM OVERVIEW

The generic overview of the proposed system is presented in Fig 1. The first step in the working of the AF system is the calibration of the eye tracker. Subsequently, a user performs reading action while the application captures visual attention pattern-related features. The key part of the overall architecture constitutes the implicit user feedback capturing process. For that purpose, a low-cost and lightweight gaze-tracking framework has been designed. It involves the usage of a single-camera image processing-based gaze-tracker and a dynamic interface for efficiently capturing gaze-related information at sentence level. We use GazeFlowAPI[5] as the eye-tracking solution.

We embed the eye tracker in a web application, WikiRead (https://wikianalysis.herokuapp.com/). We develop the application using NodeJS framework in Javascript and HTML. The application website is hosted using Heroku. Some screenshots of the web application are presented in Fig.2. This web application's primary purpose is to simplify the process of capturing Wikipedia readers' visual activity and analyze the AF-related features. On the main screen of our website, we present the welcome page of Wikipedia along with options to enter the title of the article user wishes to read and buttons to start and stop the reading session. Once a user clicks on the "Start" button, the system requests permission to share the screen content and the camera access. Camera sharing is safe for users because we do not store any images or videos of the user's face. It is required to perform eye-tracking. Once the sharing permission is granted, the control goes to the calibration[6] screen. Users are asked to follow the on-screen instructions to successfully

[4] https://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool

[5] https://github.com/szydej/GazeFlowAPI

[6] Calibration details are mentioned in the experimentation section.

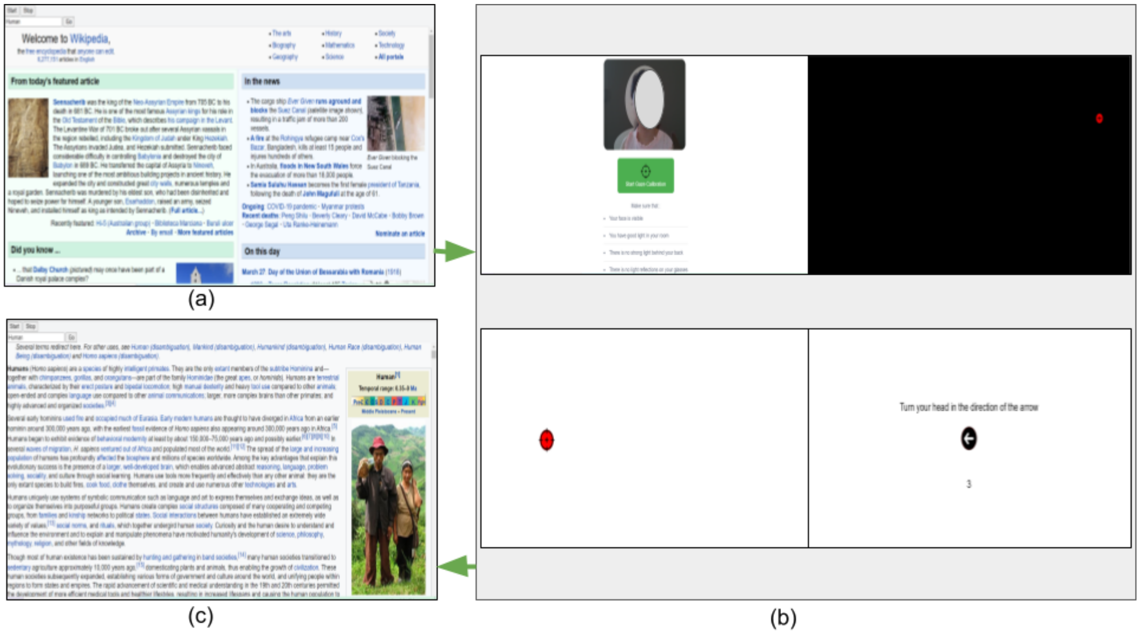


Figure 2: Various screens of the web application, WikiRead. (a) The main screen with options to enter article name and start the reading session. After which the control goes to calibration process. (b) Shows various calibration stages of the eye tracker. (c) Shows the article which user wants to read while system performs gaze tracking in background.

complete system calibration. Post calibration, the interface opens the article which the user wishes to read, and gaze tracking starts. While reading, the user can freely click on any Wikilink present in the article. Gaze data for the redirected articles are also stored accordingly. Users can finish the reading session anytime by clicking on the "Stop" button.

Following the gaze point trajectory estimation, a set of visual features are extracted only for those regions of an article that the user has observed. Then, a threshold-based feature selection procedure is applied for maintaining the most discriminative features. Table 1 shows the set of features that we capture as part of AF.

During dataset collection, we also ask the participants to manually annotate the regions of an article as per their preference level. Using the proposed framework, participants can manually select a region and rate it on a scale of 0 to 10 (10 being the most preferred region). This results in the generation of a set of automatic and manual gaze features for the sentences that have been seen by the user(s). We use MongoDB to store the dataset.

Finally, the visual AF features collected in various reading sessions are grouped as per the article title. It enables the possibility to evaluate article features using the concepts of collective intelligence. The observed benefits of social heuristics trace back to 1907, at least when Francis Galton experimented predicting the weight of an ox [15]. He showed that average estimates of median and mean were only 9 and 1 pounds, respectively, of the actual weight of 1198 pounds. This experiment proves that aggregating the crowd perspective can be helpful in cases where the size and diversity of the crowd are large. Wikipedia also receives participation from a vast and diverse crowd/users. The SOI AF (Sentences Of Interest based Attention Feedback) dataset represents crowd perspective. This knowledge can be efficiently used to analyze Wikipedia from a novel perspective.

4 ATTENTION PATTERN ANALYSIS

To capture readers' attention pattern in an article, we track their visual gaze features. These are valuable pieces of information and can be used to infer the reader's current cognitive processes [9]. Based on the gaze pattern, we extract text from the article. The below-mentioned steps are followed in sequence to track eye gaze and identify areas of interest in an article:

- Gaze tracking framework
- Attention feature selection

4.1 Gaze Tracking Framework

For each new user visiting the website (WikiRead), a randomly generated unique 20 bit code is assigned for identification. The sysId, i.e., the Identification code, is stored in the user's browser storage, and it can be deleted whenever the user deletes the storage for the site. For each unique user, a sessionID is created whenever the user starts the reading session. The sessionId is used to differentiate between various reading sessions for a single user. The website includes the English Wikipedia home page[7], in a sub-frame. The top header contains two buttons: Start and Stop and a search bar for the article title. Using Wikipedia in a sub-frame resulted in a famous web development problem called CORS[8]. This problem prohibits accessing the action events information within the sub-frame. To

[7] https://en.wikipedia.org/wiki/Main_Page
[8] https://developer.mozilla.org/en-US/docs/Web/HTTP/CORS

overcome this problem, we implemented a proxy server and routed every GET request from the Wikipedia website inside the sub-frame through the proxy server. The proxy server makes a GET request to Wikipedia, through MediaWiki API[9], and gets Wikipedia web page with headers. The proxy changes the headers of Wikipedia.org related to CORS and adds our website in the header. Same-Origin-Policy, the process displays the Wikipedia page in our website domain's sub-frame and avoids the CORS problem. However, the GET request returns a plain web page without any styles, so we apply styles explicitly to the web page every time the sub-frame reloads.

As we mentioned earlier, the sophisticated gaze-tracking systems with increased accuracy are commercially available, e.g., Tobii1, SMI2, EyeTech3, and Mirametrix4, to name the most representative ones. However, a prohibitive factor for the extensive use of such specialized equipment constitutes their significantly high cost. To satisfy the requirements of low-cost and availability, the proposed AF system follows an image processing-based approach that uses a single camera for performing gaze-tracking.

We use a non-invasive eye-tracking setup containing Logitech Webcam C922 Pro Stream and open source eye-tracker applications, named GazeFlowAPI[10]. We select this eye-tracking solution because, along with access to sufficiently accurate real-time gaze information, it also provides eye blink, and head position data. These additional pieces of information are very crucial for the consolidated evaluation of attention patterns.

In the case of non-invasive trackers (no physical contact), gaze mapping deals with inherent noise and drift problems. While mapping a single fixation to a text location may be ambiguous, the matching of groups of fixations to a chunk of objects (words/sentences) can resolve this ambiguity [32]. In the proposed method, we perform gaze point mapping at the sentence level.

For further optimization purposes, we do not store the entire reading session's screen recording. Instead, we store only the required frames for further processing. A frame is the display screen snapshot at any given time-stamp. We identify the frames by assessing generic human reading behavior. The mean minimum time to acquire the whole meaning of a word is 151 ms [42]. It has also been shown that the majority of the sentences contain 11-15 words [46]. Since the gaze feedback is performed at the sentence level, we find the average time a reader takes to perceive the meaning of a sentence. We calculate this by multiplying the data mentioned above, i.e., $151 * 15$ ms. It results in approximately 3 seconds. Therefore, we store a frame only if it is on-screen for more than 3 seconds. We store only one instance of each frame.

4.2 Attention Feature Selection

In this work, the gaze features are defined considering only the fixations since they contain more valuable information and less noise than saccades. We adopt the Dispersion-Threshold Identification (I-DT) method [53] for defining fixations. According to this definition, a fixation is considered to occur if the gaze point remains in a circular area of radius R pixels for a minimum of T ms. For the employed gaze-tracker, the following (commonly used) values were selected based on experimentation: $R = 25$ pixels and $T = 180$ ms. A fixation is denoted as $F_i(x_i; y_i; t_s; t_e), i \in [1, N]$, where N is the total number of fixations. Point $(x_i; y_i)$ corresponds to the center of the aforementioned circular area. The values of t_s, t_e are the start and end time of the fixation, respectively, where $t_e - t_s > T$. It is to note that before the fixation identification, the gaze trajectory is low-passed for noise removal (estranged points). This is performed by applying a simple mean filter separately to the horizontal and vertical gaze coordinate signals. The low pass filtering is performed according to the technique mentioned in [36].

To identify a reader's regions of interest on a given frame, we need to map their fixations on corresponding frames. On a frame, we map all the fixations with time-stamps (t_s and t_e) within the frame's screen time. After appropriately sprinkling the fixations on each frame, we plot heatmaps based on gaze-points density. The heatmap depicts the time-series gaze density distribution on a frame. Fig. 3 shows a sample frame with gaze points mapping and the corresponding heatmap of gaze points density, post noise removal process.

For each reading session, along with the gaze density heat map, we also provide a set of sentences where a user-focused while reading along with the time for which each sentence was focused. By setting an appropriate threshold on the heatmap, we extract text from each frame using the OCR tool [38]. The extracted sentences might contain noise due to the characters' misidentification by the OCR tool. After processing the sentences, we arrange them in the order they are read along with their gaze quotient. By gaze quotient, we mean the time duration (in seconds) for which a sentence is being gazed at or read. It is calculated by finding the total time duration for which gaze points are mapped within the on-screen span of the sentence's spatial location.

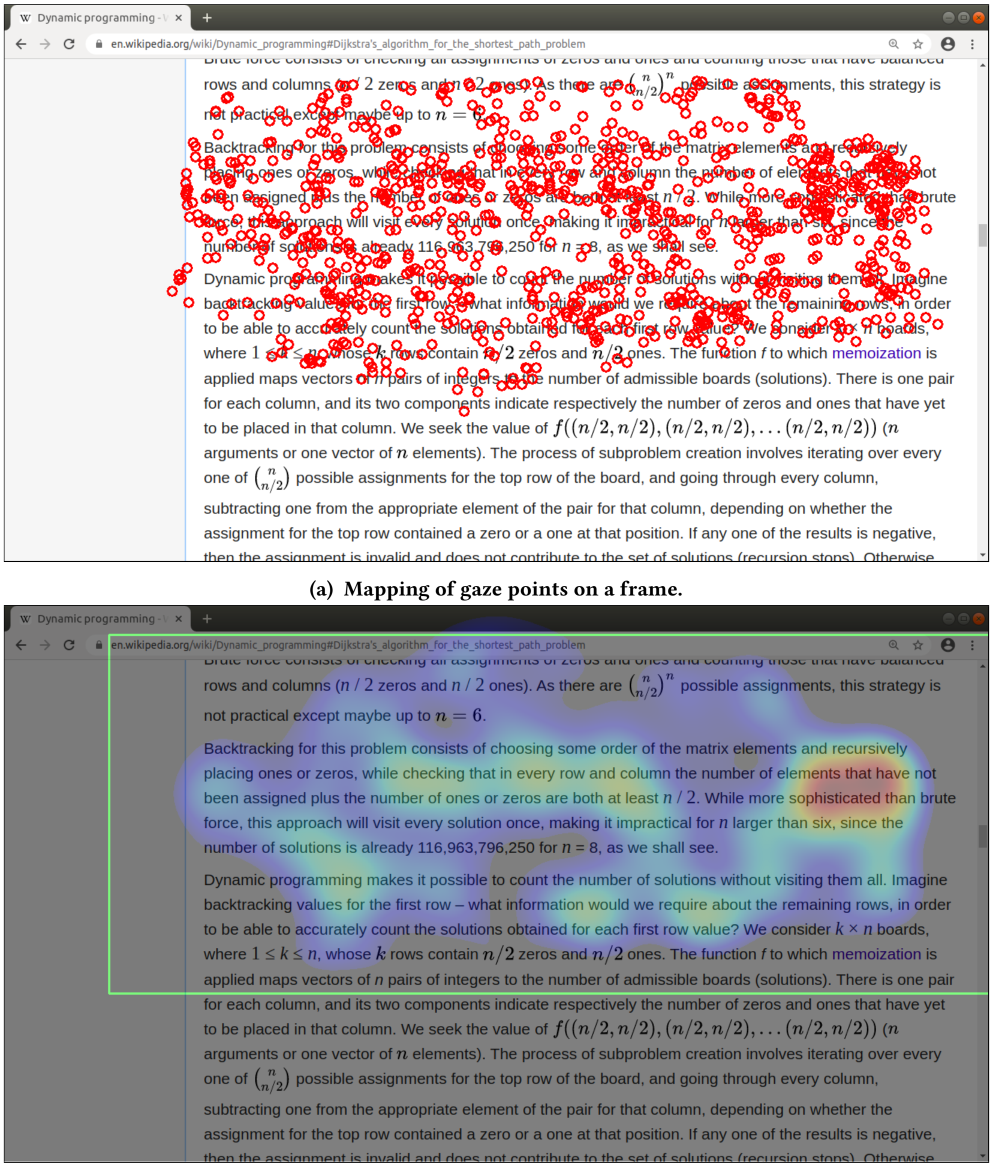
$$GQ(s) = t_s(j) - t_e(i) \tag{1}$$

In equation 1, $GQ(s)$ is the gaze quotient of sentence s. $t_s(j)$ is the start time of the first fixation on the screen span of s and $t_e(i)$ the end time of the last fixation on the screen span of s. We call the set of sentences along with their GQs as Sentences Of Interest (SOI). A user is free to click on any Wikilink. Therefore an SOI can contain sentences belonging to different articles. Later, we arrange the sentences in each SOI according to the article title.

Along with the features mentioned above, the proposed AF framework also captures some additional information listed below:

(1) Wikilink clicks: We store the time series value of Wikilink [48] click data. It can be helpful to evaluate users' navigation path or information search patterns within Wikipedia.
(2) Eye blinks: We store the time series value of readers' eye blinks. The eye blink information has been proven to be critical to assess the reader's attention level [44].
(3) Scroll events: We store the time series value of the scroll events being performed while reading. The scroll information is a good indicator of browsing strategy [26]. The GazeFlowAPI does not provide built-in support to track mouse activity. Therefore, we externally added the scroll tracking feature in the proposed AF system using PyQt5.QtWidgets package of Python.

[9] https://www.mediawiki.org/wiki/API:Main_page
[10] https://github.com/szydej/GazeFlowAPI


Dynamic programming - \
en.wikipedia.org/wiki/Dynamic_programming#Dijkstra's_algorithm_for_the_shortest_path_problem

Brute force consists of checking all assignments of zeros and ones and counting those that have balanced rows and columns (n / 2 zeros and n / 2 ones). As there are $\binom{n}{n/2}^n$ possible assignments, this strategy is not practical except maybe up to $n = 6$.

Backtracking for this problem consists of choosing some order of the matrix elements and recursively placing ones or zeros, while checking that in every row and column the number of elements that have not been assigned plus the number of ones or zeros are both at least n / 2. While more sophisticated than brute force, this approach will visit every solution once, making it impractical for n larger than six, since the number of solutions is already 116,963,796,250 for n = 8, as we shall see.

Dynamic programming makes it possible to count the number of solutions without visiting them all. Imagine backtracking values for the first row – what information would we require about the remaining rows, in order to be able to accurately count the solutions obtained for each first row value? We consider $k \times n$ boards, where $1 \leq k \leq n$, whose k rows contain $n/2$ zeros and $n/2$ ones. The function *f* to which memoization is applied maps vectors of n pairs of integers to the number of admissible boards (solutions). There is one pair for each column, and its two components indicate respectively the number of zeros and ones that have yet to be placed in that column. We seek the value of $f((n/2, n/2), (n/2, n/2), \ldots (n/2, n/2))$ (n arguments or one vector of n elements). The process of subproblem creation involves iterating over every one of $\binom{n}{n/2}$ possible assignments for the top row of the board, and going through every column, subtracting one from the appropriate element of the pair for that column, depending on whether the assignment for the top row contained a zero or a one at that position. If any one of the results is negative, then the assignment is invalid and does not contribute to the set of solutions (recursion stops). Otherwise

(a) Mapping of gaze points on a frame.

Brute force consists of checking all assignments of zeros and ones and counting those that have balanced rows and columns (n / 2 zeros and n / 2 ones). As there are $\binom{n}{n/2}^n$ possible assignments, this strategy is not practical except maybe up to $n = 6$.

Backtracking for this problem consists of choosing some order of the matrix elements and recursively placing ones or zeros, while checking that in every row and column the number of elements that have not been assigned plus the number of ones or zeros are both at least n / 2. While more sophisticated than brute force, this approach will visit every solution once, making it impractical for n larger than six, since the number of solutions is already 116,963,796,250 for n = 8, as we shall see.

Dynamic programming makes it possible to count the number of solutions without visiting them all. Imagine backtracking values for the first row – what information would we require about the remaining rows, in order to be able to accurately count the solutions obtained for each first row value? We consider $k \times n$ boards, where $1 \leq k \leq n$, whose k rows contain $n/2$ zeros and $n/2$ ones. The function *f* to which memoization is applied maps vectors of n pairs of integers to the number of admissible boards (solutions). There is one pair for each column, and its two components indicate respectively the number of zeros and ones that have yet to be placed in that column. We seek the value of $f((n/2, n/2), (n/2, n/2), \ldots (n/2, n/2))$ (n arguments or one vector of n elements). The process of subproblem creation involves iterating over every one of $\binom{n}{n/2}$ possible assignments for the top row of the board, and going through every column, subtracting one from the appropriate element of the pair for that column, depending on whether the assignment for the top row contained a zero or a one at that position. If any one of the results is negative, then the assignment is invalid and does not contribute to the set of solutions (recursion stops). Otherwise

(b) Heatmap based on gaze points density.

Figure 3: A sample gaze point distribution and gaze density heatmap on a frame after noise removal process.

Table 1 lists all the features stored in the dataset corresponding to each reading session. All the mentioned features are crucial to provide consolidated feedback from readers.

5 EXPERIMENT AND RESULTS

5.1 Experimental Setup

5.1.1 System Components. Subject's eye movements were recorded using GazeFlowAPI and a Logitech Webcam C922 Pro Stream on a VivoBook ASUSLaptop X430FN_S430FN with Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz, 1992 Mhz, 4 Cores, 8 Logical Processors, and 8GB RAM.

5.1.2 Eye Tracking Setup. The gaze tracker sampled at 30 fps, and the gaze angle was determined by the relative positions of corneal and pupil centers. The video resolution of the laptop during the test recording was 1280×720. Participants were seated 27 inches from a 14-inch flat-panel monitor that displayed the stimuli. Eye data were recorded using the proposed web application on the Google chrome browser. The embedded eye gaze tracker supports head movement. Therefore, we have not used any head/chin rest device during dataset collection.

5.1.3 Calibration Validation. Before starting the reading session, participants were instructed to watch a red dot move around the screen. The computer screen is interpreted as a coordinate system,

Table 1: AF features captured for each Wikipedia reading session performed on the proposed AF framework.

AF Feature	Description	Data Type
SysID	A unique system D for each user	20 bits
SessionID	A unique session ID for each reading session	20 bits
Article title	Title of the article being read	String
Gaze density heatmap	Heatmap based on gaze points density for each frame	Base64 encoded
SOI	Sentences of interest along with gaze quotient	List
Wikilink clicks	Time-series value of clicks on Wikilinks	List
Eye blinks	Time-series value of eye blink	List
Scroll events	Time-series value of scrolls with direction (up/down)	List

where the top left of the screen is (0, 0) and the bottom right of the screen is (1, 1). The ball's calibration movement was fixed at 16 positions in both bright and dark backgrounds sequentially with the frontal head position. Along with this, three positions each to positive and negative 45° yaw and pitch head positions. The head pose variation during dataset collection makes the gaze tracker robust to head pose variation during actual reading.

At each of the above points, the red dot paused and pulsed for ∼2s. Post calibration, if the RMSE (Root Mean Square Error) value of the actual and predicted gaze points is higher than the set threshold (0.28), we ask the participant to repeat the calibration task before starting the reading session.

5.1.4 Subjects. Seventy-two clinically normal volunteers participated in the dataset collection. The participants belong to different professional backgrounds, including engineering students, industry professionals, doctors, and professors. The ethics committee approved the study of our university. Subjects underwent informed consent procedures approved by the university, and all subjects provided written and informed consent for participation in the study. Prospective participant-observers were required to have normal or corrected visual acuity. Participants were free to wear spectacles during dataset collection[11]. The mean age of the participants was 29.7 ± 7.6 (range: 22-53). The sample was 38.5% female, and the rest male.

5.2 Evaluation of Eye Tracking Solution

The performance of the eye tracker has a significant impact on the efficiency of the overall AF approach. As a consequence, the accuracy of the employed tracker is examined in this section. Table 2 briefly outlines some of the approaches that have been evaluated for measuring image processing-based gaze-trackers' performance. We select the gaze trackers based on their suitability for the aim of this work i.e. a commodity gaze solution for mass readers. Lu et al. [29] presented an adaptive linear regression based solution for appearance-based gaze estimation. In another work, Lu et al. [28] estimated 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting. Papoutsaki et al. [37] discussed a web based gaze solution using regression analysis informed by user interactions. Computer Vision Center (CVC) released an eye tracking solution [1] which used usage of simple camera and image processing technique but this solution does not support head movements. Wang et al. [52] proposed a real-time 3D eye gaze capture with DCNN-based iris and pupil segmentation. Finally, Liu et al. [27] proposed a 3D model-based gaze tracking via iris features with a single camera and a single light Source. The restriction on light source makes this tracker less applicable in wild. The adopted GazeFlow eye tracker [2] provides high accuracy and satisfies all the requirements for this work. The comparison of GazeFlow with SMI RED 250 can be found here[12].

Table 2: Comparison of the adopted eye gaze tracker with other state-of-the-art image processing based eye trackers

Eye Tracker	Head Movement	Blink Data	Error
Lu et al. [29]	restricted	no	2.59±0.38
Lu et al. [28]	restricted	no	6.91±4.46
Papoutsaki et al. [37]	restricted	no	4.17
CVC [1]	restricted	yes	1.35±1.0
Wang et al. [52]	free	no	2.06
Liu et al. [27]	free	no	1.87±0.4
Adopted [2]	**free**	**yes**	**1.23±1.2**

For producing directly comparable evaluation results, the performance of the developed gaze-tracker was evaluated using the experimental protocols described in the works of [22]. Most of the works reported in Table 2 rely on the usage of static markers in fixed positions for evaluating the gaze tracking performance. In this way, the behavior of the gaze tracker is not adequately examined. For that purpose, a significantly more challenging and thoroughly defined experiment is proposed with the following two key characteristics: a) it can be easily reproduced b) it takes into account both the spatial accuracy (fixations and sccades) and the temporal coherence (timestamp data for Wikilink clicks, eye blinks and mouse scroll) of the tracker.

More specifically, a red dot was depicted on the screen performing the same trajectory as mentioned in the calibration process (Section 5.1.3). The participants were asked to follow the center of this red dot with their gaze. The tracker's accuracy was defined as the RMSE between the fixed points and their corresponding estimations.

Twenty individuals participated in this experiment, each performing the aforementioned task for all seven eye trackers. Regarding the specifications of the defined experiment, the monitor plane was vertically aligned, while the perpendicular vector originating from the monitor's center was maintained to approximately target the nose of the user and to also be perpendicular to the user's face plane. Additionally, the camera was placed on top of the monitor and at the center of the respective monitor's side, with the user's nose to be set to correspond approximately to the center pixel of the captured video sequence.

[11] We took precautions regarding possible glares in the spectacles of the participants during dataset collection

[12] https://www.slideshare.net/szymondeja3/raport-gaze-flow-vs-smi26092013en-1

Table 3: ROUGE-N evaluation of the proposed sentence selection technique with the manual sentence selection

Feature	Average value
P	0.717
R	0.514
F-measure	0.656

Table 4: Statistics of the collected AF dataset

Feature	value
Number of SOI_a	218
Average compression ratio of SOIs	15%
Average reading session duration	3.7 minutes

The comparative evaluation results given in Table 2 show that the developed gaze-tracker outperforms most of the state-of-art approaches and carries the additional advantage of head-movement and eye blink data. It makes this eye tracker a suitable choice for the proposed task of capturing implicit AF of readers.

At this point, it must be highlighted that the focus of this work does not include the proposal of a new gaze-tracker, whose performance needs to be accurately measured and to be superior compared to other state-of-art methods. On the contrary, the aim of this work is the proposal of a novel framework for interpreting the gaze signal and subsequently utilizing this information for realizing RF in the context of image retrieval, irrespective of the particular gaze-tracker that is used.

5.3 Evaluation of AF system

Using the above-mentioned experimental setup with seventy-two participants, we could conduct a total of 153 reading sessions. In these reading sessions, 84 unique articles were visited (including the article redirects using Wikilinks). We evaluate the performance of the proposed method against the extractive summarization task for the articles.

Precise identification of the reading pattern is required for capturing a reader's attention feedback of an article. The proposed approach identifies the sentences of interest (SOI) for each reading session and assembles them according to the article titles. To evaluate the SOI selection technique's quality, we partition all the SOIs with respect to the article titles. We call the partitioned SOIs as SOI_a. They contain sentences of a single article.

We verify the accuracy of the proposed sentence selection approach. We do so by comparing the ROUGE-N value of the sentences selected by the proposed approach with the true value i.e. manually selected sentences. To compare the manual and automated summaries, we use a modified version of ROUGE-N approach to compare the accuracy of sentence selection. ROUGE-N is an n-gram-based metric. It calculates the recall score (R), the precision score (P), and the F-measure score (F) for the text. Let S_{auto} be the set of sentences automatically selected by the proposed approach. Let S_{man} be the set of sentences in the manual summary.

$$P \triangleq \frac{Count(common\ sentences\ in\ S_{auto}\ and\ S_{man})}{|S_{man}|} \tag{2}$$

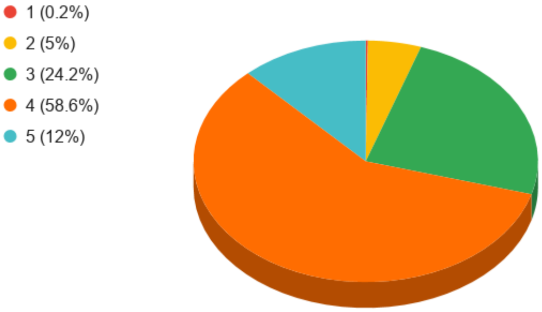


Figure 4: Distribution of users' rating for SOI selection on the scale of 0 to 5. The percentage value for 0 rating is 0% therefore it's not shown in the pie chart.

$$R \triangleq \frac{Count(common\ sentences\ in\ S_{auto}\ and\ S_{man})}{|S_{auto}|} \tag{3}$$

$$F - measure \triangleq \frac{2PR}{P + R} \tag{4}$$

The ROUGE comparison results are shown in table 3. The high value of f-measure, shows that there is high overlap between the sentences selected automatically and the sentences readers manually selected while reading the article.

We collect SOI dataset for further research purposes. Table 4 shows the values of various parameters of the collected dataset of SOIs. The count of SOIs per article indicates the number of SOIs we are able to gather to evaluate each article. Higher the number better will be precision for article analysis. The compression rate is calculated by finding the ratio between the number of bytes of an SOI and the number of bytes of the corresponding article. The average value of the compression rate for all the SOIs across articles is 15%. It implies that on average, users read only 15% of the article content. The average reading session duration is the total time a user takes to read an article, followed by the time to analyze their reading pattern. The timer starts with the calibration phase and ends with the click on the "Stop" button. The average value of completing all these tasks is only 3.7 minutes, which supports the hypothesis that readers use Wikipedia as a quick reference tool [47].

We analyze the efficiency of the SOI identification process using two evaluation techniques. The first evaluation consists of surveying all the participants and requesting them to rate their satisfaction levels with SOIs. The second evaluation involves using a formal metric (ROUGE-N), vastly used to assess text extraction procedures.

Satisfaction Survey. We gather users' perspectives regarding text extraction by conducting a survey. At the end of every reading session during dataset collection, we display on the screen the extracted SOI. We assess the user satisfaction level for SOI selection by requesting them to rate the SOI based on its similarity with the content they read in the article. The rating scale varies from 0 (low satisfaction) to 5 (high satisfaction). The source code along

Table 5: Comparison with other extractive summarization models

Model	R-L	R-1	R-2
LEAD-5	19.24	9.67	3.43
LexRank	26.23	11.89	3.12
SumBasic	22.34	14.17	5.34
SemSenSum	30.11	15.16	4.11
C SKIP	30.16	32.38	4.25
Ours	**30.92**	**37.61**	**6.26**

with the survey form is available at our Github repository[13]. The distribution of user ratings is plotted in Fig. 4. In this pie chart, the legend shows the rating with the corresponding percentage of rating. The 0 ratings are not shown in the figure because it was nil. We observe that 70% of the participants gave four or more ratings for SOI selection. The high user satisfaction ratings show the efficiency of the proposed reading pattern analysis approach.

Comparison with other extractive deep models. We compare the performance of the proposed approach with other extractive models. We include LEAD-k [45], which selects the first *k* sentences in the document as a summary. In this work, we have used LEAD-5 for comparison. LexRank [14] is a graph-based method, and it uses nodes as text units, and edges define the similarity measure. SumBasic [30] is a frequency-based sentence selection method that uses a component to re-weigh the word probabilities to minimize redundancy. The last extractive baselines are the near state-of-the-art models C SKIP [43] and SemSenSum [7]. The former exploits word embeddings' capability to leverage semantics, whereas the latter aggregates sentence semantic relation graph and graph convolution sentence embedding.

We select trained models and fine-tune them for the Wikipedia data dump[14]. This dump contains the current revision of all the English Wikipedia articles as of February 20, 2021. It does not contain talk or user pages. During fine-tune, we use a learning rate of 0.0001 with ten epochs. All models are fine-tuned on Nvidia GeForce GTX 1080 Ti GPU (60GB RAM, 12 GB dedicated graphic card, and 200 GB Hard drive space). We run all models with their best-reported parameters. Post-fine-tuning the pre-trained models, we compare these models' performance with our approach to the collected dataset of automatic and manual summaries (along with source articles for each summary). We consider three strong reference-based evaluation metrics: ROUGE 1, 2 & L. We use the recalls of ROUGE-1 (unigram), ROUGE-2 (bigram), and ROUGE-L (LCS). We use the official ROUGE script[15] (version 1.5.5) to evaluate the summarization output. In Table 5 we can see the comparison result. Our approach performs better than these models for ROUGE 1 & 2 as well as ROUGE-L metrics. The reason for the better performance of our approach for is that we are directly extracting sentences based on the user's reading pattern. Therefore, when we compare the proposed system generated summary with the manual summary, it results in high similarity.

6 LIMITATIONS & FUTURE DIRECTIONS

The goal of this work is to devise a technique to capture AF of Wikipedia readers. However, during dataset collection, we observe that some readers did not perform their natural reading. They claim that the information of their gaze data being recorded impacts their reading behavior. We counter this problem by keeping the reading interface the same as the original portal and not disturbing reading sessions with unnecessary popups of real-time feedback. As a result, most participants reported that none of the feedback conditions were distracting in any way. Therefore, the proposed gaze feedback system appears to be unobtrusive yet effective.

Due to the novelty of the proposed approach, the doors of several new future directions open. For example, we can investigate article readability based on the readers' cumulative reading pattern over an article. We can also predict the popularity of an article based on how readers refer to the article. We can potentially expand our approach into a business analytics framework/artifact that shows the analysis steps. It can also help Design Science researchers to identify the user reference behavior on the portal and thus aid them in designing the user interface. We are in talks with Wikimedia Foundation to include our application as a Wikipedia extension so that users can easily use it and researchers can obtain a vast AF dataset.

It is to note that there still exists a strong potential for further improving the user's AF prediction based on gaze data. Towards this goal, future work includes the investigation and modeling of the factors that affect the way that users see (e.g., facial expression, mood). Furthermore, their integration to the developed framework.

In the future, we also plan to extend the idea for other crowd-sourced portals, such as Stack Exchange[16], Reddit[17], and Quora[18]. We believe the proposed method can help understand users' engagement on these portals by unraveling their eye gaze information. It will help editors to collect low-cost, implicit attention feedback of portal users and may a way to a novel research direction.

7 CONCLUSION

In this paper, a novel gaze-based attention feedback approach was presented for Wikipedia users. The overall approach aimed to estimate the area of interest and reading pattern of the users and subsequently use it to derive results related to their reading behavior on the portal. A novel set of gaze features representing visual characteristics was presented for performing users' attention assessment prediction. Extensive experiments demonstrated their efficiency compared to other feedback approaches used by the Wikipedia research community. The experimental evaluation proved that the proposed approach outperforms representative reading pattern analysis approaches of the literature. Moreover, incorporating a single-camera image processing-based gaze tracker into a web application framework makes the overall system cost-efficient and portable. This study's outcomes are currently being discussed in the Wikimedia Foundation for developing specialized tools to capture readers' implicit feedback.

[13] This is an anonymous repository with lesser details. Link to the principal repository will be provided post-acceptance.

[14] https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-meta-current.xml.bz2

[15] http://www.berouge.com/

[16] https://stackexchange.com/

[17] https://www.reddit.com/

[18] https://www.quora.com/

ACKNOWLEDGMENTS

This work was funded by the assistance received from CSRI, Department of Science and Technology India via grant no. SR/CSRI/344/2016.

REFERENCES

[1] [n.d.]. CVC Eye Tracker. https://github.com/tiendan/OpenGazer Accessed: 2016.

[2] [n.d.]. GazeFlow eye tracker. https://github.com/szydej/GazeFlowAPI Accessed: 2021.

[3] B Thomas Adler, Krishnendu Chatterjee, Luca De Alfaro, Marco Faella, Ian Pye, and Vishwanath Raman. 2008. Assigning trust to Wikipedia content. In *Proceedings of the 4th International Symposium on Wikis*. ACM, 26.

[4] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[5] Antti Ajanki, David R. Hardoon, Samuel Kaski, Kai Puolamäki, and John Shawe-Taylor. 2009. Can Eyes Reveal Interest? Implicit Queries from Gaze Patterns. *User Modeling and User-Adapted Interaction* 19, 4 (2009), 307–339. https://doi.org/10.1007/s11257-009-9066-4

[6] Judd Antin and Coye Cheshire. 2010. Readers are not free-riders: reading as a form of participation on wikipedia. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 127–130.

[7] Diego Antognini and Boi Faltings. 2019. Learning to Create Sentence Semantic Relation Graphs for Multi-Document Summarization. *arXiv preprint arXiv:1909.12231* (2019).

[8] Maram Barifah and Monica Landoni. 2019. Exploring usage patterns of a large-scale digital library. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, 67–76.

[9] David Beymer and Daniel M Russell. 2005. WebGazeAnalyzer: a system for capturing and analyzing web reading behavior using eye gaze. In *CHI*. 1913–1916.

[10] Joshua E Blumenstock. 2008. Size matters: word count as a measure of quality on wikipedia. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 1095–1096.

[11] Yuru Cao, Hely Mehta, Ann E Norcross, Masahiko Taniguchi, and Jonathan S Lindsey. 2020. Analysis of Wikipedia pageviews to identify popular chemicals. In *Reporters, Markers, Dyes, Nanoparticles, and Molecular Probes for Biomedical Applications XII*, Vol. 11256. International Society for Optics and Photonics, 112560I.

[12] Quang Vinh Dang and Claudia-Lavinia Ignat. 2016. Quality assessment of wikipedia articles without feature engineering. In *JCDL'16*. ACM, 27–30.

[13] Nathan J Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *NBR* 24, 6 (2000), 581–604.

[14] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *ArXiv* abs/1109.2128 (2004).

[15] Francis Galton. 1907. Vox populi (the wisdom of crowds). *Nature* 75, 7 (1907), 450–451.

[16] Courtney Hsing and Vanessa Gennarelli. 2019. Using GitHub in the classroom predicts student learning outcomes and classroom experiences: Findings from a survey of students and teachers. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*. 672–678.

[17] Halszka Jarodzka, Thomas Balslev, Kenneth Holmqvist, Marcus Nyström, Katharina Scheiter, Peter Gerjets, and Berit Eika. 2012. Conveying clinical reasoning based on visual observation via eye-movement modelling examples. *Instructional Science* 40, 5 (2012), 813–827.

[18] Halszka Jarodzka, Katharina Scheiter, Peter Gerjets, Tamara van Gog, and Michael Dorr. 2010. How to convey perceptual skills by displaying experts' gaze data. In *Proceedings of the 31st annual conference of the cognitive science society*. Cognitive Science Society, 2920–2925.

[19] Halszka Jarodzka, Tamara Van Gog, Michael Dorr, Katharina Scheiter, and Peter Gerjets. 2013. Learning to see: Guiding students' attention via a model's eye movements fosters learning. *Learning and Instruction* 25 (2013), 62–70.

[20] Heather Knight and Reid Simmons. 2013. Estimating human interest and attention via gaze analysis. In *ICRA*. IEEE, 4350–4355.

[21] David Laniado, Riccardo Tasso, Yana Volkovich, and Andreas Kaltenbrunner. 2011. When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *AAAI*.

[22] Eui Chul Lee and Kang Ryoung Park. 2009. A Robust Eye Gaze Tracking Method Based on a Virtual Eyeball Model. *Mach. Vision Appl.* 20, 5 (July 2009), 319–337.

[23] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. 2014. Reader preferences and behavior on Wikipedia. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 88–97.

[24] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Analysis of references across Wikipedia languages. In *ICIST*. Springer, 561–573.

[25] C Li and J Bernoff. 2011. Groundswell: winning in a world transformed by social technologies, vol.

[26] Chang Liu, Jiqun Liu, and Yiming Wei. 2017. Scroll up or down? Using Wheel Activity as an Indicator of Browsing Strategy across Different Contextual Factors. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 333–336.

[27] Jiahui Liu, Jiannan Chi, Wenxue Hu, and Zhiliang Wang. 2020. 3D Model-Based Gaze Tracking Via Iris Features With a Single Camera and a Single Light Source. *IEEE Transactions on Human-Machine Systems* (2020).

[28] Feng Lu, Yue Gao, and Xiaowu Chen. 2016. Estimating 3D gaze directions using unlabeled eye images via synthetic iris appearance fitting. *IEEE Transactions on Multimedia* 18, 9 (2016), 1772–1782.

[29] Feng Lu, Yusuke Sugano, Takahiro Okabe, and Yoichi Sato. 2014. Adaptive linear regression for appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence* 36, 10 (2014), 2033–2046.

[30] Ani Nenkova and Lucy Vanderwende. [n.d.]. The impact of frequency on summarization. ([n. d.]).

[31] Blair Nonnecke, Dorine Andrews, and Jenny Preece. 2006. Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research* 6, 1 (2006), 7–20.

[32] Ayano Okoso, Kai Steven Kunze, and Koichi Kise. 2014. Implicit gaze based annotations to support second language learning. In *UbiComp*. Association for Computing Machinery, Inc, 143–146. https://doi.org/10.1145/2638728.2638783

[33] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* (2012).

[34] Pavel A Orlov and Roman Bednarik. 2016. ScreenMasker: An open-source gaze-contingent screen masking environment. *Behavior research methods* 48, 3 (2016), 1145–1153.

[35] Kai Otto, Nora Castner, David Geisler, and Enkelejda Kasneci. 2018. Development and evaluation of a gaze feedback system integrated into eyetrace. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. 1–5.

[36] Georgios Th Papadopoulos, Konstantinos C Apostolakis, and Petros Daras. 2013. Gaze-based relevance feedback for realizing region-based image retrieval. *IEEE Transactions on Multimedia* 16, 2 (2013), 440–454.

[37] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. Webgazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (New York, New York, USA) *(IJCAI'16)*. AAAI Press, 3839–3845.

[38] Chirag Patel, Atul Patel, and Dharmendra Patel. 2012. Optical character recognition by open source OCR tool tesseract: A case study. *International Journal of Computer Applications* 55, 10 (2012), 50–56.

[39] Cole S Peterson, Jonathan A Saddler, Natalie M Halavick, and Bonita Sharif. 2019. A gaze-based exploratory study on the information seeking behavior of developers on stack overflow. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[40] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2020. Quantifying Engagement with Citations on Wikipedia. In *Proceedings of The Web Conference 2020*. 2365–2376.

[41] Keith Rayner. 1978. Eye movements in reading and information processing. *Psychological bulletin* 85, 3 (1978), 618.

[42] Eyal M Reingold and Keith Rayner. 2006. Examining the word identification stages hypothesized by the EZ Reader model. *Psychological Science* 17, 9 (2006), 742–746.

[43] Gaetano Rossiello, Pierpaolo Basile, and Giovanni Semeraro. 2017. Centroid-based text summarization through compositionality of word embeddings. In *Proceedings of the MultiLing 2017 Workshop on Summarization and Summary Evaluation Across Source Types and Genres*. 12–21.

[44] Tsugunosuke Sakai, Haruya Tamaki, Yosuke Ota, Ryohei Egusa, Shigenori Inagaki, Fusako Kusunoki, Masanori Sugimoto, Hiroshi Mizoguchi, et al. 2017. EDA-based estimation of visual attention by observation of eye blink frequency. *International Journal on Smart Sensing and Intelligent Systems* 10, 2 (2017), 296–307.

[45] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).

[46] HS Sichel. 1974. On a distribution representing sentence-length in written prose. *RSS* 137, 1 (1974), 25–34.

[47] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. 2017. Why we read wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*. 1591–1600.

[48] Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012* 15 (2012).

[49] Nathan TeBlunthuis, Tilman Bayer, and Olga Vasileva. 2019. Dwelling on Wikipedia: investigating time spent by global encyclopedia readers. In *Proceedings of the 15th International Symposium on Open Collaboration*. 1–14.

[50] Dirk van der Linden, Emma Williams, Joseph Hallett, and Awais Rashid. 2020. The impact of surface features on choice of (in) secure answers by Stackoverflow

readers. *IEEE Transactions on Software Engineering* (2020).

[51] Bogdan Vasilescu, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov. 2015. Gender and tenure diversity in GitHub teams. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3789–3798.

[52] Zhiyong Wang, Jinxiang Chai, and Shihong Xia. 2019. Realtime and accurate 3D eye gaze capture with DCNN-based iris and pupil segmentation. *IEEE transactions on visualization and computer graphics* 27, 1 (2019), 190–203.

[53] Heino Widdel. 1984. Operational problems in analysing eye movements. In *Advances in psychology*. Vol. 22. Elsevier, 21–29.

[54] Dennis M Wilkinson and Bernardo A Huberman. 2007. Cooperation and quality in wikipedia. In *EC*. ACM, 157–164.

[55] Songhua Xu, Hao Jiang, and Francis Lau. 2009. User-oriented document summarization through vision-based eye-tracking. In *Proceedings of the 14th international conference on Intelligent user interfaces*. ACM, 7–16.

[56] Petro Zdebskyi, Victoria Vysotska, Roman Peleshchak, Ivan Peleshchak, Andriy Demchuk, and Maksym Krylyshyn. 2019. An Application Development for Recognizing of View in Order to Control the Mouse Pointer.. In *MoMLeT*. 55–74.

Wikipedia Edit-a-thons and Editor Experience: Lessons from a Participatory Observation

Wioletta Gluza
Aalborg University Copenhagen
Denmark
wgluza20@student.aau.dk

Izabela Anna Turaj
Aalborg University Copenhagen
Denmark
ituraj20@student.aau.dk

Florian Meier
Aalborg University Copenhagen
Denmark
fmeier@hum.aau.dk

ABSTRACT

Wikipedia is one of the most important sources of encyclopedic knowledge and among the most visited websites on the internet. As a peer-produced knowledge repository, Wikipedia is dependent on its community of contributors. A healthy contributor community and a steady stream of new editors from diverse backgrounds are especially vital for the platform's future in its endeavour of closing knowledge gaps and combating biases and a lack of diversity that Wikipedia suffers from. Edit-a-thons are social activities aiming to improve content and create new articles on Wikipedia with the purpose of recruitment and onboarding of newcomers. Although edit-a-thons have been facilitated and hosted for many years now, little is known how editors experience such events. In this paper, we study editors experience during a virtual edit-a-thon by applying an ethnomethodological perspective. We use a participatory observation to study incidents of motivation and frustration occurring during the collaborative online writing event. Moreover, we use Hofstede's 6D Model of National Culture to explore what influence culture has on participants' actions, expressed feelings and thoughts while interacting with the administrator and with each other. Our findings indicate that the type of motivational factors is very diverse and varies from general motivation to fill in knowledge gaps, in the beginning, to share good resources for citations at later stages of the edit-a-thon. However, participants also experience moments of frustration, especially concerning the usability of the editing interface and when navigating a complex bureaucracy of policies and procedures. Finally, our analysis shows that cultural idiosyncrasies can intensify the frustrating experience of social challenges.

CCS CONCEPTS

• **Human-centered computing → Wikis**; **Empirical studies in collaborative and social computing**.

KEYWORDS

Wikipedia, edit-a-thon, editor experience, Hofstede's 6D Model of National Culture

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479994

ACM Reference Format:
Wioletta Gluza, Izabela Anna Turaj, and Florian Meier. 2021. Wikipedia Edit-a-thons and Editor Experience: Lessons from a Participatory Observation. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3479986.3479994

1 INTRODUCTION

While being mostly a success story, Wikipedia and other Wikimedia projects have also been facing many challenges over time. First, as a peer-produced knowledge repository, Wikipedia is highly dependent on its online peer production community. However, Wikipedia has seen declines in the number of contributors over time and has been struggling with growing its editor base especially from the perspective of retaining new editors [4, 5]. Awareness of newcomers barriers to participation in open collaboration projects [27] have lead to the design and creation of interventions that aim to provide new editors with improved socialisation and better onboarding experience [16, 17]. While such interventions show promising results, a more sustainable and also diverse editor base including new forms of *contribution-beyond editing* are still open issues [39]. A second challenge that is closely related to the decline in editors is structural sexism and the lack of diversity among contributors (e.g. the gender gap) [11] which also results in biased and culturally less diverse Wikipedia content [1, 15, 31]. These issues have recently been summarized under the term *knowledge gaps* [39]. Only a platform that aims at higher diversity in contributorship and content will be able to achieve knowledge equity which is one of the most central goals Wikipedia is striving for [39].

In this paper, we argue that edit-a-thons are a possibility to make Wikipedia content less biased and more diverse by focusing on certain underrepresented topics as seen in e.g. both the *Women in STEM* and *Arts+Feminism* event series. At the time of writing, for 2021 the *Arts+Feminism* campaign lists 135 programs with 2309 editors who contributed almost 3000 new articles and edited another 13000 articles [34]. At the same time, edit-a-thons function as outreach events that allow for drawing in new contributors from many diverse backgrounds and countries.

Edit-a-thon is a portmanteau of *edit* and *marathon* and describes "1) a scheduled time where people edit Wikipedia together, whether offline, online, or a mix of both; 2) typically focused on a specific topic, such as science or women's history; 3) a way to give newcomers an insight into how Wikipedia works" [37]. While the idea of hosting such events was first proposed in 2004, the Wikimedia Foundation cites the British Library to be the first to use of the word “edit-a-thon” to describe their event in January 2011 [12]. Research shows that edit-a-thons support new forms of knowledge construction, which allow opportunities for the democratisation of knowledge, the diversification of Wikipedia's editor demographic and the spread of information on traditionally marginalised subjects [7]. As mentioned, such gatherings – be they held virtual or physical – also foster recruitment and integration of newcomers and increase Wikipedia literacy among participants [7]. However, many aspects of edit-a-thons still remain unknown. First, less is known about the actual retention rate of those who participated in an edit-a-thon and continue editing after the event has ended. An internal Wikimedia evaluation of edit-a-thons held in 2015 came to the conclusion that "about 52% of participants [who] identified as new users made at least one edit one month after their event, but the percentage editing dropped to 15% in the sixth months after their event" [33]. This stands in contrast to Farzan et al. findings, who estimate that only 1% of newcomers actually continue to edit Wikipedia after the edit-a-thon [3]. Second, although we can see a growing number of scholarly work on Wikipedia edit-a-thons, researchers argue that the body of research on this kind of events is still underdeveloped [30]. Previous studies highlight the complex interactions that are at play between motivations, strategies and values of facilitators and participants of edit-a-thons [12]. This implies that if we want to get a more holistic picture of the role that edit-a-thons play in the Wikimedia ecosystem — and ultimately what influences retention rates of new editors — additional research of edit-a-thons in varied contexts is necessary [12]. Finally, most studies on edit-a-thons and new editor experience on Wikipedia have been performed on English Wikipedia only [17].

We take this as a motivation to conduct and report on an ethnomethodlogically inspired participatory observation of a GLAM edit-a-thon held by a Polish editor collective. Throughout this observation, we put special attention on the incidents of motivation and frustration that occurred during the event and how they influence the participant's experience. Earlier work on Wikipedia editor experience, helped us to develop a coding scheme to account for those incidents in our analysis [26]. Furthermore, we apply Hofstede's 6D Model of National Culture [6, 8] to find a theoretical underpinning for further contextualizing of how the participants, the host and an administrator from the Wikipedia editor community interacted with each other.

Our findings indicate that the type of motivational factors is very diverse and varies from general motivation to fill in knowledge gaps, in the beginning, to share good resources for citations at later stages of the edit-a-thon. However, participants also experience moments of frustration, especially concerning the usability of the editing interface and when navigating a complex bureaucracy of policies and procedures. Finally, our analysis shows that cultural idiosyncrasies can intensify the frustrating experience of social challenges. We conclude our paper with recommendations for possible training programs and suggestions for future work.

2 RELATED WORK

In this section, we review relevant scholarly work related to our study. We present previous studies on Wikipedia edit-a-thons, academic research on Wikipedia community socialization strategies and work that covers Polish Wikipedia's editor community.

2.1 Wikipedia Editing and Edit-a-thons

A thorough and systematic review that describes all existing relevant work on Wikipedia editing and editor behaviour is close to impossible. Already in 2012, Okoli et al. reported on more than 200 publications that studied issues related to participation and collaboration on Wikipedia [20]. These studies cover a broad range of topics including motivational factors for participation (e.g. [24]) or cultural and linguistic effects on participation. For example, Pfeil et al., like we, use Hofstede's Model of National Culture to study the influence of culture on editing behaviour. They find a significant correlation, indicating that cultural background influences users behaviour on Wikipedia and the Internet in general [22]. However, none of the articles reviewed by Okoli et al. mention collaborative writing events like edit-a-thons. To the best of our knowledge, the first scientific reports on edit-a-thons can not be found before 2015.

Evans et al. report on the outcomes of the *Art+Feminsim* edit-a-thons held in 2015 over the weekend of International Women's Day (March 6-8). They argue that this event was especially successful as it fell directly into the trend of that time which uncovered the systematic bias and online harassment against women [2]. They see *Arts+Feminism* as a symbiotic way to close the gender gap in both content and participation on Wikipedia and support the idea of more feminism related edit-a-thons as a contribution to close the internet gender gap [2].

The majority of research on edit-a-thons, however, view the events with a pedagogical lens considering them as collaborative learning events [7, 21, 30]. Oliver investigated the use of edit-a-thons as a substitute for classroom assignments when teaching information literacy. Student's reflections in form of qualitative and quantitative data showed that they learned valuable lessons on researching information and writing, especially how to be critical about information sources [21]. Hood and Littlejohn, via the means of interviews, captured narrative learning stories of nine participants taking part in an edit-a-thon held at the University of Edinburgh [7]. The edit-a-thon was designed as an informal professional learning event that combines online activity with offline, in-person collaboration and interaction. Participants reported having learned valuable technical skills and increased their Wikipedia literacy. However, the reflections about biases and responsibilities in knowledge production and dissemination were rated highly among students [7]. In similar veins, Vetter and Sarraf, in their assessment of an *Arts+Feminism* edit-a-thon held at Indiana University of Pennsylvania, found that edit-a-thons are not only facilitators of Wikipedia literacy but foster critical thinking, digital literacy, and technical skills [30]. Farzan et al. studied the effect of *Art+Feminism* edit-a-thons on newcomers onboarding via the means of triangulating log data from Wikipedia and Twitter. Their results suggest that these events are very successful in attracting new members [3]. Moreover, they show that on-event support for editing during the event and social interactions can lead to higher retention rates [3]. However, according to their data only around 1% of participants can actually be called retained editors [3].

The presented studies had a strong focus on the events as a whole. One exception is the study by March and Dasgupta. March and Dasgupta interviewed 13 edit-a-thon facilitators to uncover their motivations for organizing these events and the challenges they face [12]. They discover that the personal and institutional values that inspire the events go far beyond adding content and editors to Wikipedia aim at strengthening peoples information literacy and building communities that extend beyond Wikipedia's editor community.

2.2 Wikipedia Community Socialization Stategies

Typically, collaboration on Wikipedia takes place outside of the encyclopedia's articles in the talk pages and discussion threads. It is usually also there where newcomers learn how to navigate any number of technical and organizational obstacles when they start editing and where community involvement and editor socialization takes place [14]. Already in 2013, when the decline in new editors and editor retention has been noticed for some time, additional interventions for supporting and socializing new Wikipedia editors have been introduced [16]. Morgan et al. present the Wikipedia Teahouse, a support space designed to boost overall editor retention, bridge the editor gender gap and improve general editor experience and well-being [16]. Early results proved this intervention to be effective for community diversification and new editor retention. In 2018, Morgan and Halfaker presented further evidence for the Wikipedia Teahouse success story. In a controlled study, they are able to show increased retention for both low- and high activity newcomers. Besides the Teahouse, other community socialization strategies have been proposed and tested. Such personalized socialization opportunities included mentoring programs such as *Adopt-a-User* [18] or personalized invitations to WikiProjects. However, Morgan and Halfaker suspect, that they only appealed to new editors who were already highly likely to be retained. As such, the retention potential of edit-a-thons is still not really known.

2.3 Polish Editor Community

The Polish-language edition of Wikipedia and its community of contributors has been the focus of several studies before. Many of these studies apply social network methodology to study its editor community [9, 29, 32]. Wierzbicki, Turek and Nielek, for example, propose the use of social network analysis to evaluate teams of authors on Polish Wikipedia [32]. The created network contains the dimensions trust, distrust acquaintance and knowledge. Their main goal is to measure team quality i.e. whether a group of authors contributed to a featured article. Their initial results indicated that acquaintance and trust have a positive impact on team quality. Surprisingly, distrustful behaviour turned out to be beneficial for team quality too.

Also based on article edit history, Jankowski-Lorek et al. model the process of admin elections using multidimensional behavioural networks for discovering good admin candidates [9]. They argue that the admin community isn't growing fast enough to sustain the growing needs of Wikipedia. However, they could not support their hypothesis that adminship is a closed circle. They propose a method based on editing history which could be a potential way forward to recommend new administrators.

Nielek et al. studied how elderly people can be involved in contributing to Polish Wikipedia [19]. Through a combination of task-based usability tests and in-depth interviews, the authors identified several challenges and barriers that prevented the participants from successfully contributing to Wikipedia. Among the biggest challenges they identified: an incoherent user interface, issues with the information architecture (labelling of buttons), lack of timely system feedback and a lack of guidance and support during the editing process. Apart from fixing usability and UX problems of the Mediawiki software, Nielek et al. suggest an online learning course supplemented by offline meetings to involve elderly editors in the community building process [19]. We argue that especially edit-a-thons seem like a perfect fit to realize this idea. Skorupska et al. build on Nielek et al. ideas and present the concept of a chatbot that could make Wikipedia editing and data verification more accessible, especially to elderly users [28].

To sum up, while we see significant research on editing and editor behaviour also from a community perspective and even within the Polish Wikipedia community, there has not been any study yet, that reported details on the experience of participants which could lead to further insights on how to improve editor socialization and working atmosphere at such writing events.

3 METHOD AND DATA COLLECTION

We now provide a brief description of our methodological choice and the setting in which we conducted our observation i.e. details about the edit-a-thon, the collective organizing it and its participants.

To investigate editor experience and how cultural context might influence it, we employ an ethnomethodological perspective. Our study follows the idea of a study of "talk-in-interaction within various 'institutional' or 'organisational' settings to investigate what types of interactional structures are specific to these settings" [25, p.151]. Our method can be best described as participatory observation, with two researchers actively participating in the edit-a-thon conversing with the participants when necessary. As we did not have any experience in editing Wikipedia before (nor did we have a relation to the Polish Wikipedia community) we took the role of newcomers joining the Wikipedia editing community for the first time as it is the case for many edit-a-thon participants. Data is collected in course of an edit-a-thon held by a Polish collective in November 2020.

The collective was founded, inspired by the *Art+Feminism* initiative, at the end of 2018 beginning of 2019, and is an informal group of women with backgrounds in cultural studies and art history that specialises in creating articles on female, transgender and LGBTQ artists in Polish Wikipedia. In 2020, the group had 61 active editors who created 542 new articles, uploaded 665 items to Wikimedia Commons and edited 37.6 K further articles resulting in a total of 63.9 K edits. The group and the edit-a-thons it organizes are supported by a Wikimedia Foundation Rapid Grant and Wikigrants.

The edit-a-thon lasted for four hours between 16:00 to 20:00 on 21 November 2020 and was a virtual event held via Google Hangouts. As in previous editions, the edit-a-thon followed the idea of adding new or extending existing articles of female Polish architects, photographers, illustrators etc. on Polish Wikipedia. A list of possible edits was curated by the community members in a publicly accessible Google doc before the meeting started to give newcomers the chance to work on an article without having to think about a person themselves. On that day, the group reflected different community members, including a Polish Wikipedia administrator, Wikipedia redactors, experienced editors, and new editors. In total, 16 participants joined the edit-a-thon at various stages. Almost all participants used a desktop PC to edit articles. Only one participant used her mobile device. Events like this usually do not have a strict agenda; the editors are free to decide whether they join the meeting only for some time or participate actively during the entire event. Some participants join only at the beginning of the meeting to confirm the article and ask questions to come back at the end of the event with the article ready for publication. Both the host and all participants share their screens when they encounter an issue or are willing to advise collaborators. Everyone is allowed to ask questions freely, and people who get distracted can mute the conversation. This format was beneficial to our study since we as observers could ask follow-up questions or get in-depth reflections from the participants. At the beginning of the edit-a-thon all participants were informed about the goal of the study, how data will be gathered, stored and analysed and informed consent was obtained verbally by all partakers. During the edit-a-thon, we took field notes and the complete edit-a-thon was recorded on video. The field notes and the transcript of the video recordings served as the basis of data analysis. During the data collection, the analysis process i.e. the coding of our data and the presentation of our findings, we followed the recommendations by Kawulich [10].

4 DATA ANALYSIS

In this section, we present the (1) coding scheme and (2) relevant theoretical constructs that help us with analysing and interpreting our results. Inspired by earlier work that employed ethnomethodological-inspired participatory observation, our data analysis process follows a conversation analytic approach [23]. We analyse our results in two steps: In a first step, and based on findings from earlier studies on new editor experience [26], we created a coding scheme for deductively coding the transcript on instances of motivation and frustration that participants experienced during the edit-a-thon. The New Editor Experience project focused on supporting mid-sized Wikipedias with increasing their editor base and conducted design research with new editors of South Korean and Czech Wikipedia to uncover their characteristics, behaviours and motivations as well as their key challenges [26]. As edit-a-thons are one of many aims to recruit and retain new editors, we investigated which and how many motivation incidents and challenges i.e. frustration incidents occur for participants during such a collaborative writing event.

4.1 Coding Schema

Table 1 shows the coding schema and lists all motivation and frustration incidents that got coded for. The codes and descriptions for motivations are taken from the personas published in the new editor experience project report [26, p.12]. The codes for frustration incidents and their descriptions are inspired by the summary of key challenges identified in the same study [26, p.33]. In an deductive coding process, we, whenever appropriate, applied one or multiple of these codes to a turn in the conversations among participants during the edit-a-thon.

4.2 Hofstede’s 6D Model of National Culture

New editors experience during an edit-a-thon is likely to be influenced by many context factors e.g., the size of the group, the experience of editors involved, whether it is a physical or virtual event and many more. We were interested in to what extent cultural context influences the participant’s behaviour and conversations. We turned to Hofstede’s 6D Model of National Culture to find a theoretical underpinning for how to interpret the way our participants interacted and conversed with each other. Hofstede’s model is a common framework for describing the effects of a society’s culture on the value of its members and how these affect their behavior [6]. Often, these dimensions are defined via opposing concept pairs. The first dimension, *Power Distance*, refers to the degree of how accepting less powerful members of organizations and institutions are of unequal power distributions. The second dimension, *Individualism vs. Collectivism*, describes the extent to which people feel independent as opposed to being dependent on each other as members of a larger collective. The third dimension, *Masculinity vs. Femininity* covers social role division between genders. In masculine societies, people are driven by competition, achievement and success while in feminine societies caring for others and a focus on the quality of life is more important. The fourth dimension *Uncertainty Avoidance* covers how threatened a society feels by ambiguous or unknown situations and which measures (e.g. habits and rituals) are in place to deal with the anxiety of an unpredictable future. The fifths dimension, *Long Term Orientation*, describes how welcoming a society is to societal change and how strong it maintains links with its own past. The sixth and final dimension, *Indulgence vs. Restraint*, covers how strongly people in a society follow their desire and impulses and give room to socialisation with others and leisure time as opposed to duty being the normal state of being. Figure 1 shows a bar chart that compares the three countries China, Denmark and Poland on the aforementioned six dimensions, and gives a good impression of how different cultural orientations in societies can be according to that scale.

Polish society can be characterized as an individualist society in which members mostly take care of themselves and their immediate families [8]. In individualist societies offence causes a loss of self-esteem and guilt. While highly individualistic, Polish society also has a strong need for a clear hierarchy. Following [8] high numbers in both *Individualism vs. Collectivism* and *Power Distance* causes tension in a culture. Furthermore, Poland is a masculine country in which people live in order to work and conflicts are resolved by fighting them out. "Managers are expected to be decisive and assertive" [8]. Moreover, Polish people have a strong tendency to avoid uncertainty which makes them follow rigid codes of beliefs and rules. This leads to a hard-working culture with high precision. However, innovation may be resisted. Regarding their *Long Term Orientation* Polish culture can be considered as normative (not pragmatic) with a sense for traditions. Finally, Polish society is a restrained society. Restrained societies do not put much focus on leisure time. They "feel that indulging themselves is somewhat wrong" [8]. We will use these characteristics of Polish society to interpret the role of main actors (i.e. the host and the administrator) and participants behaviour during the edit-a-thon.

Incident	Code	Description
Motivation	Promoting myself or my work	Edits are motivated by leveraging Wikipedia's reach for promoting their work or themselves
	Achieving a goal or completing a task	Edits are motivated by smaller but tasks e.g. adding a reference or a picture
	Gaining experience or knowledge	Edits and actions are motivated by the drive to learn something new and be part of the offline community
	Fixing bugs or filling knowledge gaps	Desire to fix smaller 'common sense' errors or correcting false claims
	Sharing my knowledge	Desire to share topical knowledge or knowledge about good resources
	Contribute to the changes in society	Edits are motivated by a strong passion for addressing biases or imbalances on often controversial topics
Frustration	Unintuitive user interface	Usability issues: New editors struggle with using the visual and source editor or have problems with the UI in general
	Unclear Wikipedia's procedures and policies	Conceptual challenges: there is a lack of clarity on when and how Wikipedia's procedures and policies apply
	Unavailability of Wikipedia as a formal institution	Wikipedia seems formal, academic and authoritative, which makes new editors feel that their edits have to be perfect
	Communication challenges and issues with receiving support	New editors struggle with finding the help they need and in communicating with other users

Table 1: Coding schema inspired by [26] and applied in the data analysis process.

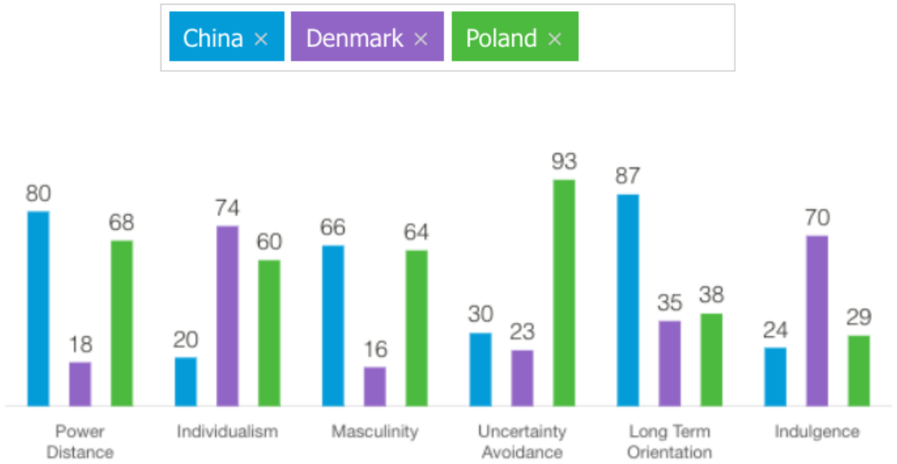
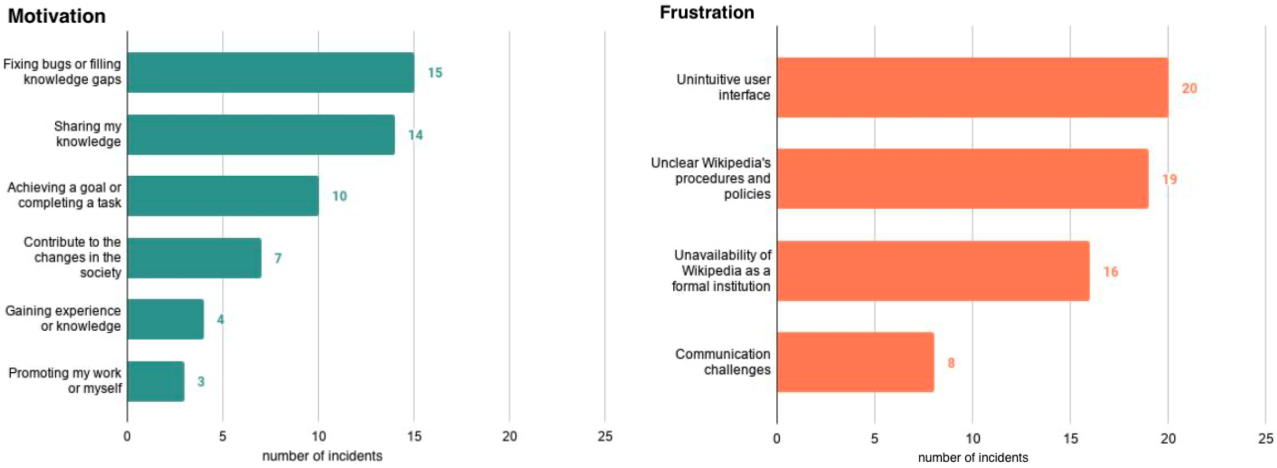


Figure 1: Comparison of three countries, China, Denmark and Poland, along their scores for Hofstede's 6 Dimensions of National Culture. This Figure was created via a tool accessible at [8].

5 FINDINGS

During the four-hour meeting, the group accomplished to publish eleven new articles and extend two existing ones. Observing this process and the surrounding discussions, we recorded a total of 53 motivation incidents and 63 frustration incidents. These findings are broken down in Figure 2.

Fixing bugs, closing knowledge gaps (n=15) and sharing one's own knowledge (n=14) were the most frequent motivational aspects that occurred during the session (see Figure 2). Due to the fact that the edit-a-thon followed a common goal, the number of instances in which participants wanted to promote their work or themselves was low (n=3). The type and frequency of motivational aspects vary depending on the phase of the edit-a-thon and the specific tasks at hand. In the beginning, participants expressed excitement for writing biographical notes about their favourite female artists and spreading knowledge about them in society via a publicly accessible Wikipedia article. At a later stage, editors were looking forward to sharing knowledge about good sources for references and strategies and how to find them. Finally, towards the end, when participants had their initial drafts published as articles, they got motivated to begin new articles out of the contentment of having achieved the task of publishing despite many hurdles and frustrating incidents. When looking at it from a time series perspective, many motivational incidents were followed by a frustration incident. The most common frustration incidents recorded were usability problems and



(a) Motivation incidents and their frequency

(b) Frustration incidents and their frequency

Figure 2

UX issues i.e. problems resulting from an unintuitive user interface (n=20). On multiple occasions during the edit-a-thon participants complained about not being able to find specific functions/buttons in the publishing process e.g. for starting to edit a page, for moving an article from the draft section over to the publishing section or which functions to use to finally publish an article. This was both the case for the visual and the source editor. Another step in the publishing process that caused confusion among participants was the step of assigning articles to categories as many editors were of the impression that this is done automatically. Participants also discussed the UX flaws of experimental features like the translation tool, which none of the participants was recommending to use [13]. Moreover, the administrator condemned behavior in which articles only get translated from English to Polish. This is in line with previous observations, that especially among senior editors and administrators, the idea of rather not having an entry for an article then a 'bad' translation of an English article is widespread.

In some cases, frustration incidents were overlapping i.e. coded with multiple codes. In the case of article categorization, for example, confusion was caused by the lack of official rules regarding which and how many categories should an article be assigned to (Unclear Wikipedia procedures and policies (n=19)). Some of these cases were resolved by the host sharing her screen and sharing good technical practices or standard procedures like copying categories from similar types of articles or linking to other already existing articles. The host suggested creating red links which from her point of view hinted towards knowledge gaps and possibilities to grow smaller language editions. This was also linked to a motivation incident as one editor remarked that spotting red links motivate her to contribute more to Wikipedia and close such gaps. Another example is finding the ideal structure for a biography entry. Here, the host would advise to find a 'best article' and copy paste it's source code.

Finally, instances or episodes which reflected the unavailability of Wikipedia as a formal institution (n=16) are mostly related to situations in which the administrator decisively referred to official Wikipedia guidelines, which prevented participants from using a certain reference or writing an article about a specific person. Details on episodes like these are described further below.

The interactions and conversations among the edit-a-thon participants result in qualitative data which has the potential to provide insights into the culture of the Polish editor community. Using Hofstede's model we try to shed light on the influence of national characteristics on contribution to Wikipedia and editors experience. As mentioned in section 4.2 Poland scores high in the dimension *Power Distance* and the 6d model describes it as a hierarchical society in which people accept a hierarchical structure and top-down orders within an organization. The behaviours recorded during our observation demonstrate this characteristic. As a powerful person, the Wikipedia administrator expressed strong opinions about Wikipedia rules and policies. On multiple occasions throughout the edit-a-thon the administrator enforced rules regarding verifiability and citations, copyright of material or notability of a person very strictly which caused frustration among editors who already struggled with conceptual challenges of that kind when contributing to Wikipedia.

The second dimension *Individualism vs. Collectivism* characterises Polish society as an individualistic society, which contradicts the hierarchical order, resulting in tensions within a society that scores high on both dimensions. During the edit-a-thon, this cultural factor may have lead experienced editors to challenge administrators' rules which created tensions. During the observation, these incidents became evident when the host of the meeting or one of the more experienced editors did not agree with the policies brought forward by the administrator and shared opinions about Wikipedia being a free encyclopedia. Especially, when the administrator brought forward aspects of notability i.e. whether an artist was respected enough to have a Wikipedia article, the host and senior editors hinted towards the new encyclopedic rules for young modern female artists who do not fall into the typical rules on Wikipedia's notability criteria [36]. During the discussion, many new editors mentioned that with rigid policies in place closing knowledge gaps and overcoming biases in Wikipedia is difficult. They argued that societal structures i.e. keeping women away from award ceremonies etc. make it impossible to find the necessary sources that would document their significance and that they are *worth of notice.* While this might not be so much the case nowadays, it is a problem for historical figures. However, it is mostly articles about women that get deleted by administrators due to an alleged lack of notability.

All in all, unclear rules and a strict application of policy by the administrator often lead to this kind of articles being blocked from publications during the edit-a-thon. The Polish culture, according to the Hofstede's model, is a masculine culture, with a high drive for avoiding uncertainty and a clear long term orientation [8]. This means that society is characterised by a normative culture suspicious of change in which conflicts are being solved by fights rather than compromise. This became obvious in most discussions between editors, the host and the administrator as the administrator did not really engage in dialogue but instead referred to the official rules blocking any real discussion. Everything that seemed unorthodox or not compliant with the official rules appeared to make him feel insecure which is probably why he insisted on the strict application of policies. This seems also counter-intuitive as one of Wikipedia's main rules is: *Ignore all rules (IAR)* [38]. Moreover, he acted decisive and assertive — as expected from managers in masculine cultures — and the tone of voice in which this dialogue was lead by the administrator was rather harsh. This caused a certain tension in the group which consisted mainly of younger female editors. Following the discussion with the administrator, multiple editors complained that even on discussion pages in Polish Wikipedia, prolonged discussions are rather the norm rather than the exception and are often impossible to solve via achieving consensus. Nevertheless, one has to mention that the administrators' help has also been perceived as helpful for example when he hinted towards documentation.

Finally, the low score for indulgence for Polish culture was expressed among Wikipedia editors with the pessimism in bringing *actual* change. This became evident in two cases: first one participant mentioned that articles about women are generally rejected by administrators although they would have well-researched references and sources. Second, it became evident in the discussion at the end of the observation that participants showed uncertainty about their skills and competencies to change the rules for modern artists' publications on Wikipedia. The contributors discussed if it was more meaningful to instead create a list of possible changes and new articles and pass this list on to administrators as they did not truly believe in having the skills to do this themselves.

6 DISCUSSION AND CONCLUSION

Wikipedia needs more Wikipedians and those should ideally come from diverse backgrounds. Only by recruiting new editors with various differing ethnic, racial, geographic and cultural backgrounds holding various perspectives e.g. political beliefs or sexual orientations, the platform can achieve it's goals of bridging knowledge gaps and creating knowledge equity [39]. We believe that edit-a-thons have the potential to be part of solving these problems. Despite being around for many years now, systematic investigations of how editors experience these collaborative writings events are missing to a large degree. In this paper, we reported on a participatory observation conducted during a *Arts+Feminism* edit-a-thon held by a Polish editor community in November 2020. To learn more about the participant's experience, we recorded, transcribed and coded incidents of motivation and frustration that occurred during the collaborative writing event, as well as the interaction between the participants, the host and the administrator. We used Hofstede's 6D Model of National Culture as a theoretical basis to explain the behaviour of all people involved. Our findings support the previous results by research on mid-size Wikipedias exposing communication issues and unintuitive interface as problematic aspects of editor experience [26]. While those problems have been known for several years now, especially the usability issues haven't been fixed yet.

However, in our findings, it became yet again evident that the greatest challenges new editors face are not technological but conceptual and cultural. Thus, this is a level where cultural identity measurable via Hofstede's index has a huge influence on editor experience. As new editors struggle with understanding Wikipedia policies, finding help and receiving feedback, the administrator's role and how she/he conveys policy rules, engages in discussions and frames feedback is essential. However, especially experienced editors and administrators struggle with providing personal, constructive feedback and e.g. explaining the rationals of policy rules to new editors. Previous studies found multiple explanations for this [26]. First, as experienced editors and administrators develop more advanced levels of their skills they become removed from the new editor experience and lose their competence to relate to new editors and their problems. Second, experienced editors and administrators feel that their efforts of mentoring are wasted if new editor investment and retention rate is low [26]. It is especially in these moments that issues expressed and experienced during the edit-a-thon, such as gender bias or the organization's authoritative character, may be intensified by the characteristics of Polish culture described with Hofstede's metric [6]. Polish society is characterised as a strong male-focused society with a high *Power Distance* which we interpret as what lead the administrator to a strict application of Wikipedia policies and the unwillingness to engage in discussions and thus also the lack of contextualization and clear description of rationale behind the policies. However, this left many of the new editors frustrated and culminated to a point where they themselves were in doubt whether they could bring any change to Wikipedia and whether it wouldn't be more effective to leave possible changes to Wikipedia for the administrator to do. From our point of view, this attitude is yet again related to Polish culture which is restraining people from fulfilling existing needs.

Our study is limited to the degree that we did not conduct multiple observations in different Wikipedia communities and cultures and performed a comparative analysis of the influence of culture on new editor experience during edit-a-thons. Thus we can not be entirely sure whether the editors experience and the general climate during the edit-a-thon are a result of Polish culture, or whether it was mostly influenced by the administrators personality. Moreover, the strong gender imbalance, i.e. most editors being female and the administrator being male, could have also played a role.

However, the goal of this study is not to generalize across multiple communities and events. Our aim is to contribute to research on editor experience during edit-a-thons and raise awareness of the fact that culture can have a huge impact on newcomers experience during edit-a-thons. We also believe that this sets the opportunity to improve training material that is offered for edit-a-thon hosting. The Wikimedia foundation already provides excellent material on how to host edit-a-thons and other editing events in three modules [35]. One option would be to add additional material for coaching experienced editors and administrators on how to provide meaningful feedback and mentorship to new editors while keeping potential influences of cultural dimensions in mind. This material could be tailored towards different regions and cultures depending on how they score on the 6D scale.

With this article we try to highlight the necessity of future research on edit-a-thons and what influence a communities' culture has on editor experience and editor retention. Future work should perform a comparative analysis of edit-a-thons hosted in countries with differing positions on Hoftede's scale. Finally, there is a need to investigate the influence of edit-a-thons on newcomer retention rate in greater detail in general.

REFERENCES

[1] Ewa S. Callahan and Susan C. Herring. 2011. Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology* 62, 10 (2011), 1899–1915. https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.21577

[2] Siân Evans, Jacqueline Mabey, and Michael Mandiberg. 2015. Editing for Equality: The Outcomes of the Art+Feminism Wikipedia Edit-a-thons. *Art Documentation: Journal of the Art Libraries Society of North America* 34, 2 (2015), 194–203. https://doi.org/10.1086/683380

[3] Rosta Farzan, Saiph Savage, and Claudia Flores Saviaga. 2016. Bring on Board New Enthusiasts! A Case Study of Impact of Wikipedia Art + Feminism Edit-A-Thon Events on Newcomers. In *Social Informatics*, Emma Spiro and Yong-Yeol Ahn (Eds.). Springer International Publishing, Cham, 24–40.

[4] Ryan Faulkner, Steven Walling, and Maryana Pinchuk. 2012. Etiquette in Wikipedia: Weening New Editors into Productive Ones. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (Linz, Austria) *(WikiSym '12)*. Association for Computing Machinery, New York, NY, USA, Article 5, 4 pages. https://doi.org/10.1145/2462932.2462939

[5] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline. *American Behavioral Scientist* 57, 5 (2013), 664–688. https://doi.org/10.1177/0002764212469365

[6] Geert Hofstede. 2001. *Culture's Consequences : Comparing Values, Behaviors, Institutions, and Organizations across Nations*. Sage Publications.

[7] Nina Hood and Allison Littlejohn. 2018. Hacking History: Redressing Gender Inequities on Wikipedia Through an Editathon. *The International Review of Research in Open and Distributed Learning* 19, 5 (2018). http://www.irrodl.org/index.php/irrodl/article/view/3549

[8] Hofstede Insights. 2021. Country Comparison. https://www.hofstede-insights.com/country-comparison/poland/. [Online; accessed 26-January-2021].

[9] Michal Jankowski-Lorek, Lukasz Ostrowski, Piotr Turek, and Adam Wierzbicki. 2013. Modeling Wikipedia admin elections using multidimensional behavioral social networks. *Social Network Analysis and Mining* 3, - (2013), 787–801. https://doi.org/10.1007/s13278-012-0092-6

[10] Barbara B. Kawulich. 2005. Participant Observation as a Data Collection Method. *Forum: Qualitative Social Research* 6, 2 (2005). http://nbn-resolving.de/urn:nbn:de:0114-fqs0502430

[11] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. 2011. WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (Mountain View, California) *(WikiSym '11)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/2038558.2038560

[12] Laura March and Sayamindu Dasgupta. 2020. Wikipedia Edit-a-Thons as Sites of Public Pedagogy. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 100 (Oct. 2020). https://doi.org/10.1145/3415171

[13] MediaWiki. 2021. Content translation — MediaWiki. https://www.mediawiki.org/w/index.php?title=Content_translation&oldid=4372284. [Online; accessed 26-January-2021].

[14] Amanda Menking and Jon Rosenberg. 2021. WP:NOT, WP:NPOV, and Other Stories Wikipedia Tells Us: A Feminist Critique of Wikipedia's Epistemology. *Science, Technology, & Human Values* 46, 3 (2021), 455–479. https://doi.org/10.1177/0162243920924783

[15] Marc Miquel-Ribé and David Laniado. 2018. Wikipedia Culture Gap: Quantifying Content Imbalances Across 40 Language Editions. *Frontiers in Physics* 6 (2018), 54. https://www.frontiersin.org/article/10.3389/fphy.2018.00054

[16] Jonathan T. Morgan, Siko Bouterse, Heather Walls, and Sarah Stierch. 2013. Tea and Sympathy: Crafting Positive New User Experiences on Wikipedia. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. ACM, New York, NY, USA, 839–848. https://doi.org/10.1145/2441776.2441871

[17] Jonathan T. Morgan and Aaron Halfaker. 2018. Evaluating the Impact of the Wikipedia Teahouse on Newcomer Socialization and Retention. In *Proceedings of the 14th International Symposium on Open Collaboration* (Paris, France) *(OpenSym '18)*. ACM, New York, NY, USA, Article 20. https://doi.org/10.1145/3233391.3233544

[18] David R. Musicant, Yuqing Ren, James A. Johnson, and John Riedl. 2011. Mentoring in Wikipedia: A Clash of Cultures. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration* (Mountain View, California) *(WikiSym '11)*. ACM, New York, NY, USA, 173–182. https://doi.org/10.1145/2038558.2038586

[19] Radoslaw Nielek, Marta Lutostańska, Wiesław Kopeć, and Adam Wierzbicki. 2017. Turned 70? It is Time to Start Editing Wikipedia. In *Proceedings of the International Conference on Web Intelligence* (Leipzig, Germany) *(WI '17)*. Association for Computing Machinery, New York, NY, USA, 899–906. https://doi.org/10.1145/3106426.3106539

[20] Chitu Okoli, Mohamad Mehdi, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2012. The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia. *SSRN* October 24, 0 (2012), 1–138. https://ssrn.com/abstract=2021326

[21] J.T. Oliver. 2015. One-shot Wikipedia: an edit-sprint toward information literacy. *Reference Services Review* 43, 1 (2015), 81–97. https://doi.org/10.1108/RSR-10-2014-0043

[22] Ulrike Pfeil, Panayiotis Zaphiris, and Chee Siang Ang. 2006. Cultural Differences in Collaborative Authoring of Wikipedia. *Journal of Computer-Mediated Communication* 12, 1 (2006), 88–113. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2006.00316.x

[23] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. ACM, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[24] Pattarawan Prasarnphanich and Christian Wagner. 2009. The Role of Wiki Technology and Altruism in Collaborative Knowledge Creation. *Journal of Computer Information Systems* 49, 4 (2009), 33–41. https://www.tandfonline.com/doi/abs/10.1080/08874417.2009.11645338

[25] G. Psathas. 1995. "Talk and social structure" and "studies of work". *Human Studies* 18, - (1995), 139–155. https://doi.org/10.1007/BF01323207

[26] Reboot. 2017. *New Editor Experience. Summary of Findings from Korean and Czech Wikipedia.* Technical Report. Wikimedia Foundation. https://upload.wikimedia.org/wikipedia/commons/0/08/New_Editor_Experiences_summary_of_findings%2C_August_2017.pdf

[27] Aaron Shaw and Eszter Hargittai. 2018. The Pipeline of Online Participation Inequalities: The Case of Wikipedia Editing. *Journal of Communication* 68, 1 (2018), 143–168. https://doi.org/10.1093/joc/jqx003

[28] Kinga Skorupska, Kamil Warpechowski, Radoslaw Nielek, and Wieslaw Kopec. 2020. Conversational Crowdsourcing for Older Adults: a Wikipedia Chatbot Concept. In *Proceedings of 18th European Conference on Computer-Supported Cooperative Work, Siegen, Germany, June 14-17, 2020.* European Society for Socially Embedded Technologies (EUSSET). https://doi.org/10.18420/ecscw2020_ep05

[29] Piotr Turek, Justyna Spychała, Adam Wierzbicki, and Piotr Gackowski. 2011. Social Mechanism of Granting Trust Basing on Polish Wikipedia Requests for Adminship. In *Social Informatics*, Anwitaman Datta, Stuart Shulman, Baihua Zheng, Shou-De Lin, Aixin Sun, and Ee-Peng Lim (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 212–225. https://doi.org/10.1007/978-3-642-24704-0_25

[30] Matthew A. Vetter and Krista Speicher Sarraf. 2020. Assessing the Art + feminism Edit-a-thon for Wikipedia literacy, learning outcomes, and critical thinking. *Interactive Learning Environments* 0, 0 (2020), 1–13. https://doi.org/10.1080/10494820.2020.1805772

[31] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *Proceedings of the International AAAI Conference on Web and Social Media.* https://aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10585

[32] Adam Wierzbicki, Piotr Turek, and Radoslaw Nielek. 2010. Learning about Team Collaboration from Wikipedia Edit History. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration* (Gdansk, Poland) *(WikiSym '10)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/1832772.1832806

[33] Wikimedia. 2015. *Learning and Evaluation/Evaluation reports/2015/Editathons.* https://meta.wikimedia.org/wiki/Learning_and_Evaluation/Evaluation_reports/2015/Editathons

[34] Wikimedia. 2021. *Campagin: Arts+Feminism 2021.* https://outreachdashboard.wmflabs.org/campaigns/artfeminism_2021/programs

[35] Wikimedia. 2021. *Running Editathons and other Editing Events.* https://outreachdashboard.wmflabs.org/training/editathons

[36] Wikipedia. 2021. *Wikipedia: Notability (people).* https://en.wikipedia.org/wiki/Wikipedia:Notability_(people)

[37] Wikipedia. 2021. *Wikipedia:How to run an edit-a-thon.* https://en.wikipedia.org/w/index.php?title=Wikipedia:How_to_run_an_edit-a-thon&oldid=1010183187

[38] Wikipedia contributors. 2021. Ignore all rules - Wikipedia The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Ignore_all_rules&oldid=1000097319. [Online; accessed 26-January-2021].

[39] Leila Zia, Isaac Johnson, B M, Jonathan Morgan, Miriam Redi, Diego Saez-Trumper, and Dario Taraborelli. 2019. Knowledge Gaps – Wikimedia Research 2030. https://figshare.com/articles/journal_contribution/Knowledge_Gaps_Wikimedia_Research_2030/7698245/1

Extracting and Visualizing User Engagement on Wikipedia Talk Pages

Carlin MacKenzie
The University of Edinburgh
Edinburgh, United Kingdom
s1724780@ed.ac.uk

John R. Hott
The University of Virginia
Charlottesville, United States of America
jrhott@virginia.edu

ABSTRACT

As Wikipedia has grown in popularity, it is important to investigate its diverse user community and collaborative editorial base. Although all user data, from traffic to user edits, are available for download under a free and open license, it is difficult to work with this data due to its scale.

In this paper, we demonstrate how consumer hardware can be used to create a local database of Wikipedia's full edit history from their public XML data dumps. Using this database, we create and present the first visualizations of how editing on talk pages differs between user groups. Our visualizations demonstrate that low quality edits are primarily performed by IP users, rather than blocked users, and that overall engagement with talk pages has plateaued over the last 10 years across all user groups. Finally, we investigate the feasibility of classifying blocked users using this dataset as an example of future research directions. However, we demonstrate the difficulty of this task and find that additional data or a more advanced model would be needed to classify them, as our approach didn't provide sufficient information to do this.

We anticipate that our visualizations and data extraction process are of interest to the community and will provide researchers with the tools needed to use Wikipedia's valuable data when resources are limited.

CCS CONCEPTS

• **Information systems → Data extraction and integration**; *Wikis*; • **Human-centered computing → Information visualization**.

KEYWORDS

Wikipedia, classification, data visualization, talk pages, data extraction

ACM Reference Format:
Carlin MacKenzie and John R. Hott. 2021. Extracting and Visualizing User Engagement on Wikipedia Talk Pages. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3479986.3479995

1 INTRODUCTION

Twenty years after its release, Wikipedia needs no introduction. Started as an experiment in anonymous, public collaboration, it is now the largest and most popular reference work on the Internet [1]. Additionally, its talk pages feature some of the most high-quality debate on the Internet which provides fertile ground for research [4][21]. However, too little attention has been paid to the engagement of users on talk pages as a vehicle for understanding and approximating user and group activity across Wikipedia as a whole [15]. We extract user interactions on talk pages throughout the entire edit history, along with group membership among the users, in order to visualize and depict engagement patterns across and among the users and groups. We anticipate these visualizations will provide (i) additional insights on how editors engage with each other on these pages and (ii) demonstrate whether talk page interactions provide a comprehensive enough depiction of the overall diversity in editor engagement to classify users into membership in one of these groups.

Analyzing Wikipedia data presented some challenges, due to several limiting factors. Like Wikipedia itself, documentation about accessing the data is community generated, with little top-down guidance of best practices or official tools. We found multiple publicly curated lists of tools[1], each providing a combination of maintained and abandoned projects. None of the tools found provided an efficient mechanism for procuring user engagement data and edit history.

Although the data is made as public as possible, the direct exports of the database tables are not published. Instead, the full edit history of all pages is released as a set of $\approx$ 600 archives, each of which extract to $\approx$ 50 GB XML files. Other options are provided to access individual files or limited revisions, however none address the need to consume the entire edit history. For example, alternative versions of the XML data files, denoted as "multistream" [19] and compressed using bzip2 [16] with multiple bzip2 streams per file, are provided to allow indexing and extracting a particular page without the need to decompress the entire file. Additionally, if a smaller section of Wikipedia is of interest, such as a category or a set of pages, the Special:Export[2] tool can be used. Unfortunately this tool is limited to the 1,000 oldest or newest revisions.

Since our work seeks to depict interactions among users and groups across all history, using Special:Export or multistream data

[1] https://meta.wikimedia.org/wiki/Datasets#Tools_to_extract_data_from_Wikipedia:, https://meta.wikimedia.org/wiki/Data_dumps/Other_tools, https://meta.wikimedia.org/wiki/Data_dumps/Tools_for_importing, https://web.archive.org/web/20191218101830/http://wikipapers.referata.com/wiki/List_of_tools
[2] https://en.wikipedia.org/wiki/Special:Export

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479995

files are insufficient. Therefore, we focus on processing the published data dumps consisting of all Wikipedia edits[3].

We extract and analyze edits on talk pages from the published data dumps to depict user engagement on Wikipedia. We anticipate that there is a larger diversity in editor engagement with these pages compared with other pages, such as those in main space, as users are making requests and having discussions with other users [15]. We separate users into several distinct groups, such as special users, blocked users, IP users and bots. This provides a framing for visualization and classification.

More specifically, in this study, we aim to explore the diversity of editor engagement on talk pages through a discussion of the following research questions:

RQ 1. How can the Wikipedia dataset be made more accessible, in terms of reduction in size or computation time, for research using consumer hardware?

RQ 2. Do visualizations of the data and metadata extracted when addressing RQ1 provide insights into differentiating Wikipedia user groups?

RQ 3. Does the metadata extracted while addressing RQ1 alone provide enough context about user edits to classify blocked users?

In response to RQ1, we present an open-source tool to create a database of edits for any Wikipedia namespace (Section 3). It downloads and partitions each dump before extracting the diff and metadata (which we refer to as *features*) of each edit. Next, regarding RQ2, we present the first visualizations depicting how different user groups interact with Wikipedia talk pages (Section 4). We then investigate RQ3, by demonstrating a future area of research with this data (Section 5). We attempt to classify blocked users using a linear classifier on the calculated features, rather than the edits themselves; while providing a provoking depiction of the group, it was unable to adequately classify the individual users. In the last section, we discuss our findings and outline future research directions.

2 RELATED WORK

Multiple studies have emerged in an attempt to understand user engagement on Wikipedia. Recent work has focused on classifying actors, detecting vandalism, analyzing the content of talk pages, and extracting a social network of users. In terms of visualizing talk page data, focus is mostly on user interaction, specifically regarding edit wars [3][13] or deletion discussions [18].

Rawat et al. [14] attempted to accurately classify abusive actors. They acquired their data by scraping user contributions from Wikipedia and applying machine learning to this data set. Their model provided an 84% accuracy, however the data set they used was very small. Our research instead provides a vehicle for expanding this dataset to all Wikipedia edits.

In the field of vandalism detection, Javanmardi [8] provides a high performing and fast model. They used a data set of Wikipedia edits which were manually classified to be spam or not spam. They created a classifier with 66 features which had an accuracy of 95.5% Area Under Curve (AUC) on the test set. To create the high performing model, they used the Lasso technique which resulted in 27 features and 95% AUC.

Schneider et al. [15] discuss the articles' talk pages. They aimed to classify the diversity in these pages and created thirteen such categories. They explored how users could signal which category their edit belonged to for aggregation purposes. They found that the most controversial articles have relatively short talk pages due to repetitive arguments in which neither side convinces the other. In contrast, Martinez-Ortuno et al. [11] looked at users' talk pages and how this is related to user activity. Compared to article talk pages, user talk pages can be thought of as the user's profile where people can thank or ask questions of the user. This is therefore a good predictor of a user's standing in the community. The researchers did not find a direct correlation between negative messages and decreased edits, but did present a model that could be used to predict user edit activity.

From a user standpoint we can look at the social networks of Wikipedia. Massa [12] found that extracting a network-based dataset could be approached in three ways, each of them flawed. Manual extraction is the most reliable but very time consuming and would not find edits which were reverted. Scraping talk pages faced many challenges such as custom signatures. Finally, they used the Wikipedia XML dumps. This was the most accurate for finding user's edits but could not verify to whom users were replying.

3 DATA ACQUISITION

In order to process the large amount of data in the compressed XML data dumps, and address our first research question RQ1, we created a tool NSDB (Namespace Database) [9] to extract all edits from a user-provided namespace. Succinctly, our tool temporarily downloads, splits, and extracts edit diffs and user-interaction features to create a complete SQL database of edits for any Wikipedia namespace, which is specified on the command line. Based on user parameters, it limits the amount of temporary storage and number of cores used.

3.1 Namespace Database

To ensure the tool is of use to future researchers, NSDB was developed in Python, since it is a popular and stable language for Data Science. For data storage, the tool utilizes either a MySQL or MariaDB database, since they are both robust and open-source. The tool has been released as open-source on GitHub[4], including thorough documentation.

Downloading the XML dumps can take significant time due to the location of any given mirror, high load, throttling, traffic, etc. Consequently, a speed test is performed to each mirror before the dump is downloaded. As each dump is independent of the others, we can easily parallelize parsing. To maximize parallelization, and localize errors, we split each dump into several partitions. This functionality is performed by a helper tool, `splitwiki`, which splits each dump at page boundaries into N partitions—123 per file on average—preserving the header with each one, to create a valid

[3] https://dumps.wikimedia.org

[4] https://github.com/carlinmack/NamespaceDatabase

Table 1: Features extracted from edits

(1) comment contains “copyedit”
(2) comment contains “Personal Life”
(3) comment length
(4) ratio of special char’s in comment
(5) # inserted internal links
(6) # inserted external links
(7) len. of longest inserted word
(8) len. of longest inserted char. sequence
(9) ratio of inserted capitalization
(10) ratio of inserted digits
(11) ratio of inserted special char’s
(12) ratio of inserted whitespace
(13) ratio of # pronouns to # words
(14) number of deleted words
(15) if the article was blanked
(16) if they inserted vulgarity
(17) if they are a special user

smaller dump. This means that the storage requirement of the partitioned dumps is slightly greater than the singular file, however they allow for more efficient parsing and fine-grained error handling.

Each partition is parsed by `parse`. The parser streams each page in the partition, extracts features and inserts them into the database. MediaWiki’s mwxml tool [6] is used for efficiently streaming the XML, allowing it to be iterated in Python with low memory overhead. Since each edit in the dump is stored as the resulting full page, NSDB must extract the diff between the current and previous edit to determine exactly which text was modified. The wdiff [10] tool was used to provide a word-level difference between the edits, extracting both the added and/or deleted text. If the page is in the selected namespace, this difference is then stored in the database to provide a more granular user-based edit history. From these texts, measurements about the text–*features*–are calculated and stored alongside the edit. When deciding which features to extract, we followed the methodology of Javanmardi [8]. In total we extracted 17 out of 27 of their features, listed in Table 1.

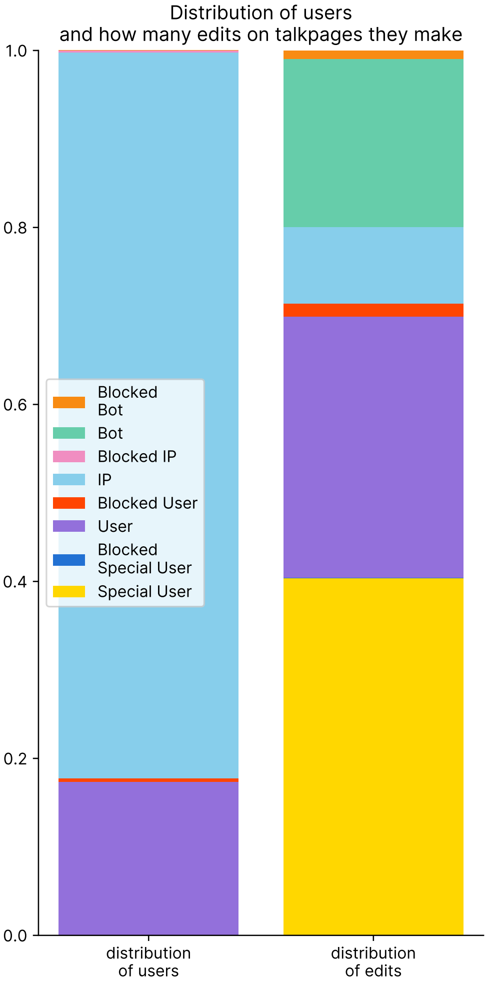


Figure 1: Distribution of users across groups and the total number of edits made by each group. While IP users constitute the bulk of talk page editors, they edit only a small fraction of pages.

Features without implementation details were not implemented.

If the page is not in the selected namespace, NSDB calculates basic statistics on the editor, including number of edits and reverted edits, which is stored with the user information. This means that for the target namespace we insert into the database for every revision, whereas in the non-target namespace we insert once for each user that makes an edit. Finally, we add information about the page to the database.

In the database, there are four main tables:

- **edit**, which stores each edit for the target namespace, including added and deleted text as well as calculated feature values;

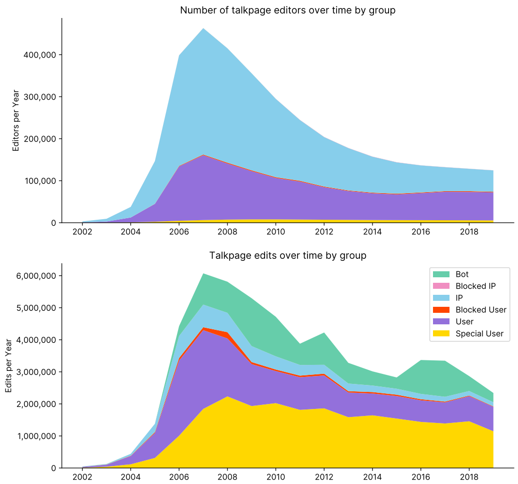


Figure 2: Number of active editors on talk pages across time by user group (above) compared with the total number of talk page edits per year by user group (below).

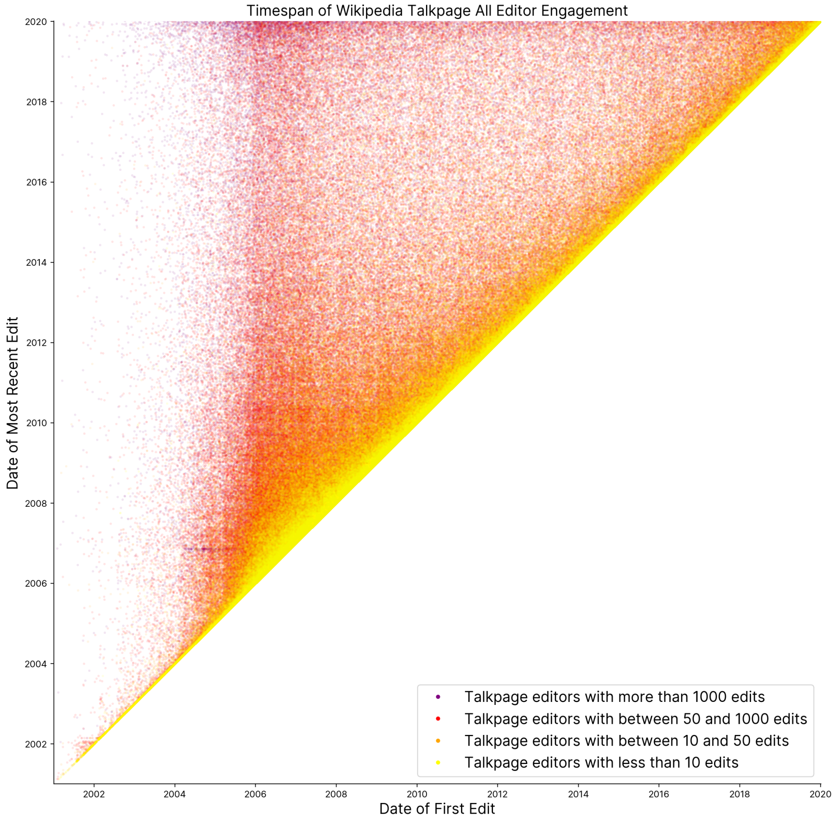


Figure 3: User engagement with talk pages plotted as their first and most recent edits. Users who do not engage long with talk pages are depicted along the diagonal. Each user's number of edits is shown by the hue of their point.

- **page**, which stores information about pages such as the title, namespace, and number of revisions;
- **user**, which stores user information and overall metrics for them, including number of edits and reverted edits across all namespaces;
- and **partition**, which provides bookkeeping to keep track of NSDB processing tasks.

Error handling was important as we expected high variation in content. Errors were logged to a file unless parsing was stopped, in which case they were logged to the database. Partitions which failed were manually restarted and the cause of failure was investigated and fixed.

Coordinating these processes is `nsdb`, the main application. It downloads a list of current XML dumps and starts the process of downloading, splitting, and processing in parallel. It allows the user to specify the location of the temporary data dump and partition files, the maximum space it should use, and the number of cores it should keep available for other processes on the system.

Depending on the timeline needed for NSDB's parsing, NSDB may be run independently or in combination with a Slurm Workload Manager [20] and Python multiprocessing to increase parallelization. Slurm may be used to distribute `nsdb` on several nodes, while `nsdb` itself uses multiprocessing to create several processes of `splitwiki` and `parse` on the same node.

3.2 Performance

All processing was performed on namespace 1, the article talk space. This is a namespace that only 10% of users edit and receives far less editing than main space. We parsed the publicly available April 1, 2020 dump release, which consisted of 660 individual XML files and resulted in 75,355 partitions. The resulting database was 83GB in total, with the edit table being the bulk of the database, at 68GB.

3.2.1 Consumer Hardware. NSDB was developed using a laptop with 8GB RAM, 8 cores, and limited hard drive space. Due to space constraints, only one dump could be processed at a time. However, we found that NSDB was able to process approximately 2–3 partitions per minute in this environment.

3.2.2 Ingest at Scale. In order to speed up processing, we used external computing resources in addition to some occasional processing locally. We used 6 compute nodes, with 14 cores on each. One thread was reserved for `nsdb`, 3 for `splitwiki` and 10 for `parse`. Using our final diagnostics, we averaged 7 partitions per minute for 132 hours. On average, we processed 5 dumps per hour (660 dumps total). We saw at least a 100 times reduction in database size compared with dump size.

3.2.3 Consumer Trade-offs. Based on our initial benchmarks, we predict that the database could be created using a single personal

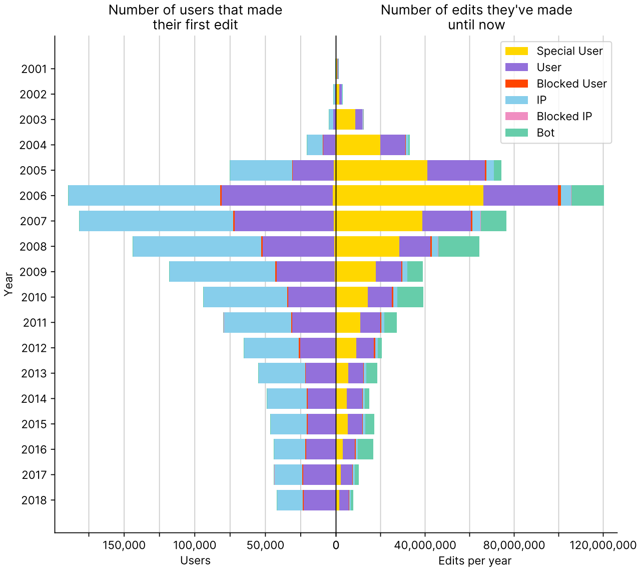


Figure 4: Comparing the number of users who made their first talk page edit in a given year with how many edits those same users make in the future. The users who began in 2006 have made approximately 120M edits since their first interaction; a majority of those edits were made by special users even though they constitute only a small fraction (1.2%) of users who began that year.

computer with several weeks of constant processing. Disk space is the main bottleneck on consumer hardware, as there needs to be 100GB of space free per dump due to the brief doubling of space required during partitioning. Utilizing external disks or focusing either only on a sample of the dumps or a category of pages may significantly increase compute time and reduce the time needed to complete the parsing.

3.3 Extensions and Research Accessibility

NSDB provides a turn-key approach to extracting, simplifying, and reducing the size of the Wikipedia XML data dumps into a manageable MySQL database, confirming our first research question, RQ1. By exploiting the inherent parallelism of the download, extraction, and feature computation tasks, we have created a solution that scales from space-limited personal computers to server-level hardware. However, further benchmarking is still required on mid-level hardware, such as current consumer desktops, to determine the processing time on those systems. Additionally, we did not investigate the size and time taken to parse the main article namespace. However, an estimate from the diagnostics we performed would suggest it would be at least 10 times slower and produce a database that is at least 10 times larger.

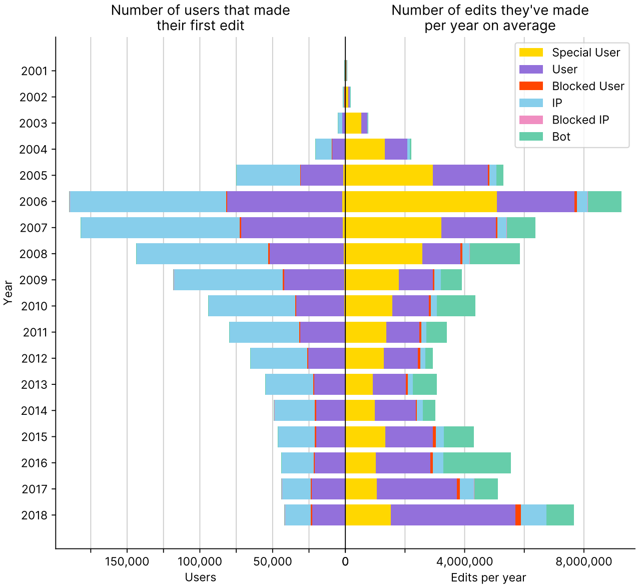


Figure 5: Comparing the number of users who made their first talk page edit in a given year with how many edits they make per year on average. The users who began in 2018 and 2006 appear to be among the most active groups.

4 VISUALIZING TALK PAGE DATA

After creating the database of edits, we investigated user groups and user engagement over the history of Wikipedia. Our second research question, RQ2, asks if visualizations of this data can provide insights into differentiating user groups and their behaviors over time. To begin addressing this question, we chose six groups of interest to focus on in our visualizations and analysis. From largest to smallest, they are:

- IP users, which are not logged in;
- Users, which are logged in;
- Blocked users, which are users that have been blocked;
- Blocked IP users, which are users that are not logged in but the IP address they are using has been blocked;
- Special users, which are users that have been given privileges (such as being able to protect pages or rollback many edits at one time); and
- Bots, which are accounts which are operated automatically.

Figure 1 shows the breakdown of these user groups in the database. We can see that there exists a reverse correlation between group size and number of talk page edits made. The most substantial group of editors is the IP users with 70% of the total, however they make only 9% of the edits. One of the smallest groups, special users, creates the most edits. Even though blocked users are a small subsection of the overall user population, they make a surprisingly large number of edits.

As we begin to incorporate time into the visualizations, Figure 2 temporally extends the bar charts of both user composition and number of edits from Figure 1. These area graphs depict the number

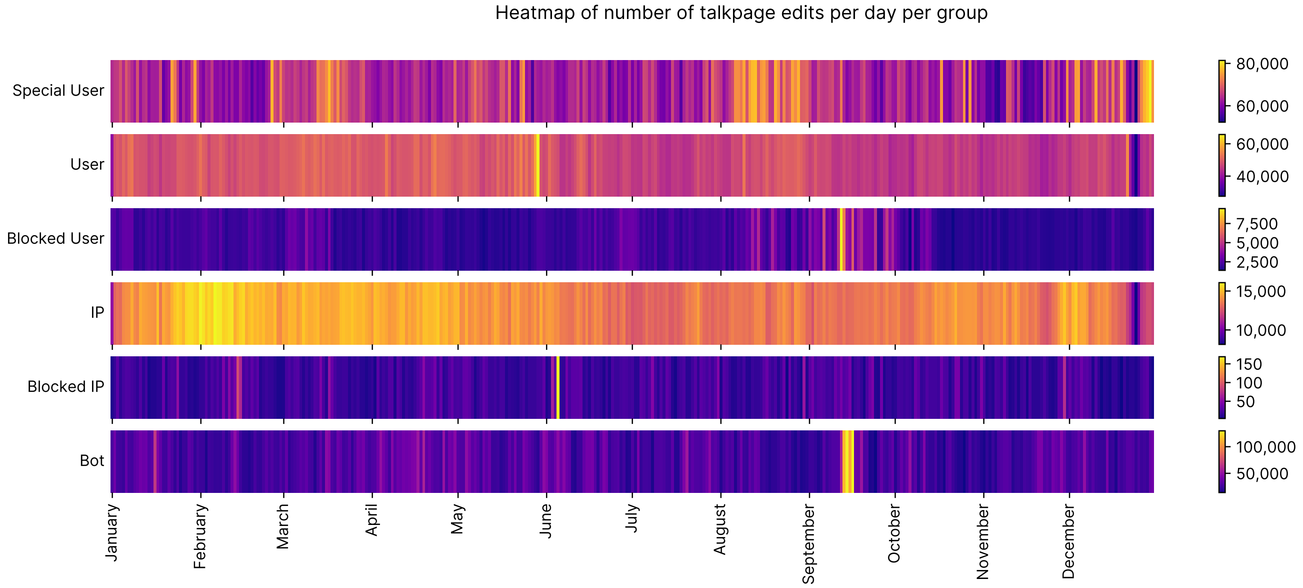


Figure 6: Heatmap showing the activity of user groups per day of the year.

of editors active each year along with their group classification and the number of edits made overall each year. This demonstrates that talk page activity peaked in 2007 and the number of engaged users has steadily plateaued. The number of edits made has decreased in all groups, besides bots, but the activity of special users appears to be decreasing at a slower rate which shows that they are remaining engaged.

4.1 Capturing Time and History

Figure 3, inspired by the work of Bégin et al. [2], was created to investigate the duration that users are active on Wikipedia's talk pages. The first and last edits of each user are plotted, with the total number of edits they have made encoded by the hue of the point. In this visualization, we see many users with relatively few edits along the diagonal, denoting that they made an edit but did not stay engaged with the site. A dense group of long-time editors, with large numbers of edits through 2020, created their accounts in 2006–2008 and have continued editing since. However, it appears that users who began editing after 2008 do not engage with the site as long. Some posit that the decline in the following years was due to the automated edit filters that were set up, which turned people away from Wikipedia [7][17].

To continue investigating the vertical bands of extended engagement from users who started editing around 2006, we created a second novel visualization of similar data in Figure 4. It presents a portion of the data from the previous visualization to connect the number of users who started editing in each year with the number of edits they have created since.

Specifically, the left-hand side of the plot shows how many users made their first edit per year. Directly connected to that bar on the right-hand side, we show how many edits their accounts have made until now. For example, with uniform data, such as if 100 accounts were created each year and each account made 100 edits per year, the left-hand side would show a constant 100 for each year. However, the right-hand side would decrease since the accounts in each successive year would have less time to edit. We do not observe this trend in the actual data. The editors who began in 2006 have made the most edits, surpassing those who came before them, and appearing to have an out-sized role compared with future editors as well. We surmise that this highlights the very active group of core users, which is also evident in Figure 3. The height on the right in this figure is decreasing, but not uniformly as expected; there is a plateau from 2013 to 2016, which demonstrates an increase in editing by some new users.

In order to understand these trends, our visualization includes the breakdown of user groups. Even though there are a few special users who started editing in 2006, they have contributed the most edits of all user groups. Specifically, these 2,344 users have made 66.2M of the 120.3M edits of all users who started in 2006. Conversely, in the plateau of 2013 to 2016, while the number of user and special user edits decreased, the amount of bot edits increased.

Another interesting trend noted in this visualization is the composition of user groups compared with their edit behaviors. A bulk of the first edits, shown on the left-hand side, are edits by IP addresses–specifically, users who are not logged in. They consisted of a majority of first edits from 2005 through 2015. Since 2016, a positive trend has emerged showing that more editors are creating accounts to make their first edits rather than editing via IP. In all cases, the IP users do not have longevity on the site, since even though they are a majority of first edits each year, they are a small fraction of the continued future edits of that year.

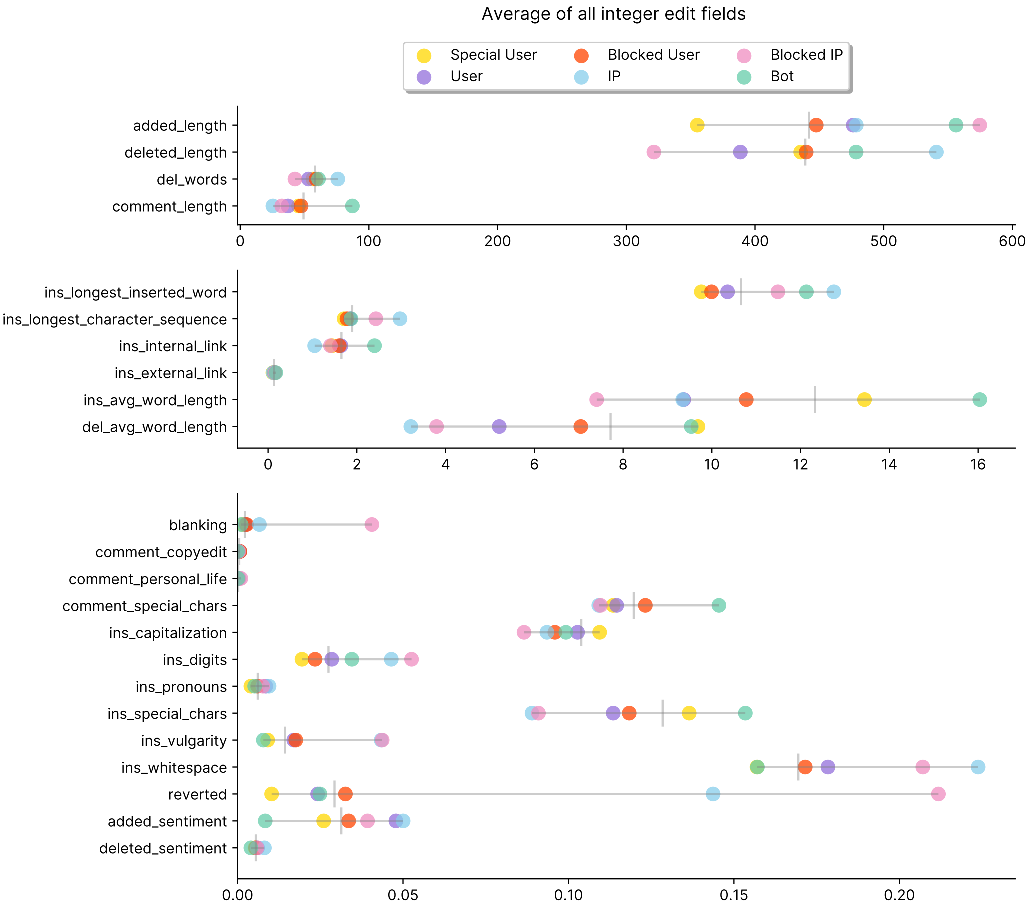


Figure 7: Average value of each feature taken from talk page edits, displayed per group.

Figure 5 provides a slightly different view, replacing the total number of future edits for each year on the right-hand side with the average edits per year. This visualization removes the confounding problem of decreasing future time inherent in the previous figure. For our simple example of 100 users making 100 edits per year, this view would make both sides constant: 100 users starting each year with each yearly set of users making 10,000 edits per year on average. Through this diagram, we see the out-sized role of the special users in 2006; they make more edits per year on average than each group of users make per year from 2009 through 2016. In fact, it highlights the special user engagement across all the years, since they make a substantial number of edits even though they are only a small fraction of the new users in any given year. In 2013, users began to overtake special users by average number of edits per year. Users who started editing in 2017 and 2018 make the most edits compared with other types of users that started the same year. This may denote a shift towards more engagement and edits by users rather than their special user counterparts, or perhaps these users are a part of the next cohort to be given permissions.

Conversely, the IP users' edits average out as we look backwards in time through the figure. Indeed, Figure 4 shows that IP users do not make a large number of edits overall relative to other users each year; that is, they appear to make only a handful of edits per IP address. Therefore, this phenomenon is expected, as the IP users will account for a proportional number of edits in the year they started, but very few in subsequent years.

Next, we take a closer look at the time that users in each group edit and engage with talk pages. Specifically, the days of the year that different groups of users edit provides another important metric of each group's engagement. Figure 6 shows a heatmap across time of how many edits on average each group makes per day of the year. Each group has a drastically different editing pattern, however they share some similarities as well. We see that there is a consistent drop in editing across all the groups over the holidays in late December, with the largest drop being in IP users. There is a peak in blocked activity in September, which coincides with the start of the new academic year. Plots which are mostly blue or yellow have high variance between high and low days of editing respectfully. We see that special users and users have low variance implying that they frequently edit regardless of the day of the year.

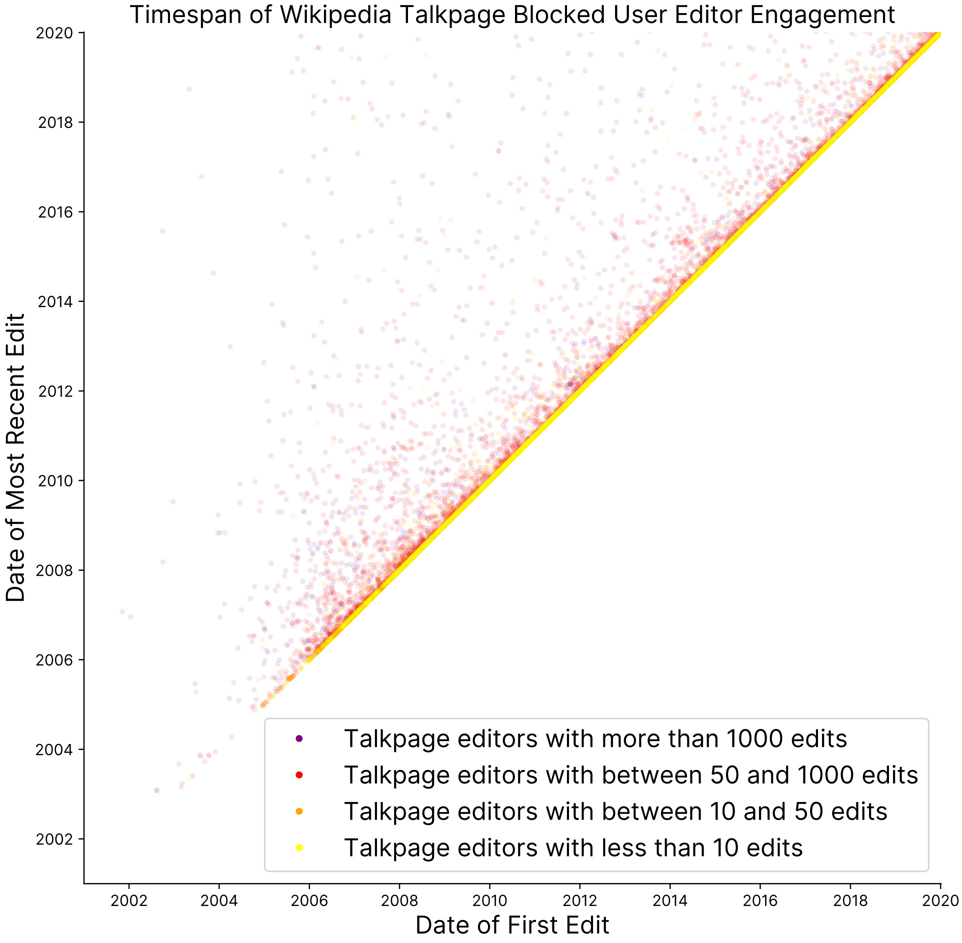


Figure 8: Blocked users tend to engage only for short periods of time, as depicted by the heavy-weight diagonal.

4.2 Capturing Group Edit Features

Moving away from temporal visualizations, we averaged the calculated features for each group's talk page edits. Figure 7 is a connected dot plot which shows each group's average value for every feature. The range is shown with a horizontal grey bar for clarity and the average value is depicted by a vertical line. This plot depicts how each user type engages with talk pages, evoking the possibility of a rubric to measure users against these group trends. For example, the plot shows that on average blocked IP users and bots added the most content per edit, whereas special users added the least. In contrast, blocked IP users delete the least per edit on average while non-blocked IP users and bots delete more per edit. Some distributions are more difficult to interpret, like inserted capitalization, than others, like blanking—where blocked IPs and IPs tend to remove the entire contents of talk pages the most.

From this chart we can tell that edits that we think of as spam, such as high vulgarity and high reversions, are mainly made by IP and blocked IP users. Blocked users are sometimes higher in these regards, but in general are very similar to users. Also, special users, or users with privileges, aren't regularly found on the extremes like we might expect, or when they are, it is hard to justify why that would be the case. The only feature that they are on the extreme of, which follows intuition, is they are reverted the least. As they edit the most, they generally pull the all-user average towards them. Finally, it seems that on average all user groups add content with positive sentiment. This suggests that people are generally positive and nice when talking to each other on talk pages.

4.3 Discussion

Across our visualizations of user edit history and edit features, trends emerge among the different groups of users. Special users and bots tend to make the bulk of the edits compared with general users and IP users. Blocked users and bots tend to edit more in mid-September, while special users, users, and IP users tend to have a constant stream of edits all year. Likewise, blocked IP users tend to be reverted more, insert shorter words, and make longer edits overall. We can affirm that these visualizations help differentiate these groups and highlight their differences, confirming RQ2.

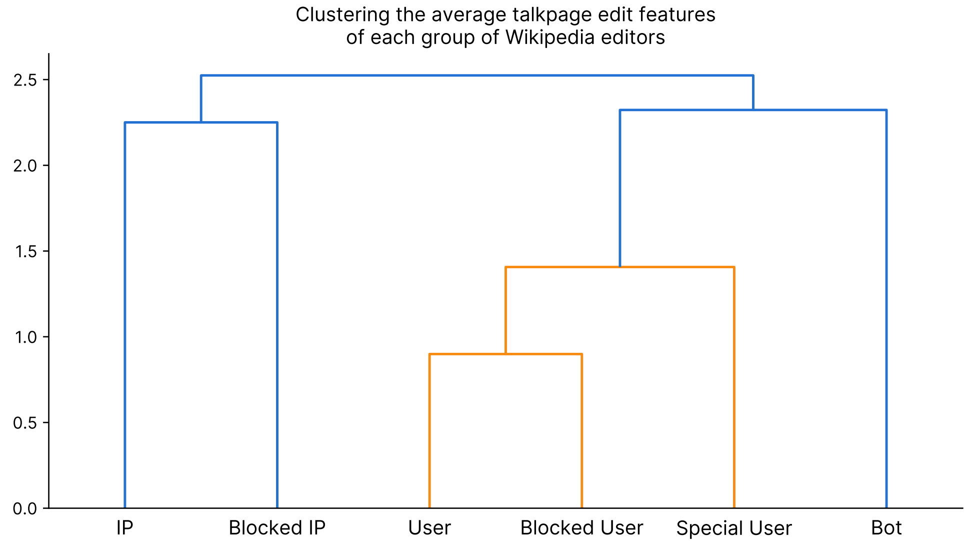


Figure 9: A dendrogram, using the centroid distance function, showing the similarity of talk page editing features of different groups. Users and blocked users are shown to be the most similar groups.

5 CASE STUDY

Building on the trends apparent in our visualizations, we introduce a case study to address RQ3, in which we investigate whether blocked users can be classified based upon features calculated from their edits alone.

First, to get a signature of their edits over time, we isolated blocked users from Figure 3 to create Figure 8. We see that blocked users tend to edit only for a short time and make few edits overall. This may be due to being blocked from editing quickly, or it might be indicative of their edit habits. To further investigate, we attempted to linearly classify blocked users using the average feature values for the group as shown in Figure 7. By combining the average feature values of blocked users compared to other groups, and the timelines typical of these users, we anticipated that a signature would emerge that would identify users that should be blocked. Users that have been blocked on average add less than their unblocked counterparts, while deleting more text, use smaller longest words, include more special characters with less white space, and use text with a lower computed sentiment. Unfortunately, we found that the average values for each of these features were not indicative of the individual members of the group—there was a high variance among the members that confounded the ability of the classifier to correctly identify blocked users.

We verified the difficulty in classification using clustering as depicted in the dendrogram shown in Figure 9. A dendrogram [5] is a visual representation of a clustering algorithm, where at each step the closest points in N-dimensional space are clustered together and depicted as siblings on the visualization. As a side-effect, since nodes which are clustered earlier are more similar, their connections appear lower in the visualization. The normalized average features for each user group were used as the 23-dimensional input and then clustered. We note that users and blocked users are shown to be the most similar, on average, and therefore harder to separate and classify. This is also the case for IP and blocked IP users, however the distance between them is greater.

We then hypothesized that, as users would not have edited after they were blocked, their most recent edits would potentially be the cause of their blocking. Thus, we instead focused on the feature values of the users' last five edits and how that differs from their averages. In order to ensure data quality, we only considered users that had made 10 or more talk page edits; this resulted in a total user population of $223,722$, consisting of $4,314$ blocked users. Figure 10 depicts the averages of features for the blocked users of this smaller population against their overall feature averages. We see a higher amount of adding and deleting of text in blocked users' last edits, along with shorter average word lengths, more vulgarity and white space, and a significant spike in reversions. However, even though these trends are apparent when comparing the change in blocked user behavior shortly before their last edit, we see similar behaviors among other user groups as well. Figure 11 plots the difference between the average feature value and the average value of the last five edits for all user groups. Even though the blocked user's edit patterns change, they do not appear to evidence drastically different behavior compared with other groups. Therefore, this difference is not indicative enough to use as a signature to classify a user as one who should be blocked, and we can answer RQ3 in the negative.

6 FINDINGS

It is clear how important 2006 and 2007 were for gaining high impact and dedicated users that continue editing to this day. Overall, the number of talk page editors has plateaued over the last decade, but unfortunately the total number of edits per year has declined since. However, we can see promising trends of new, productive editors joining.

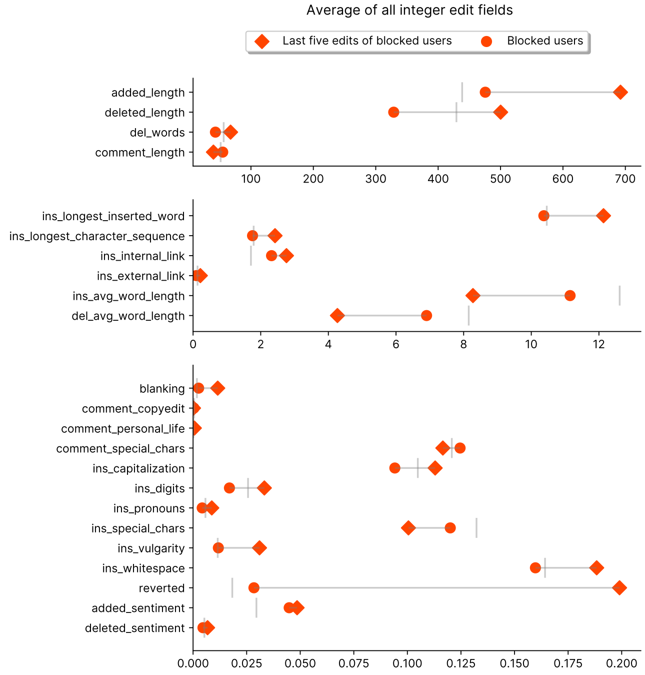


Figure 10: Comparing the features of blocked user's last five edits versus their average. Blocked user's edits are reverted more before they are blocked.

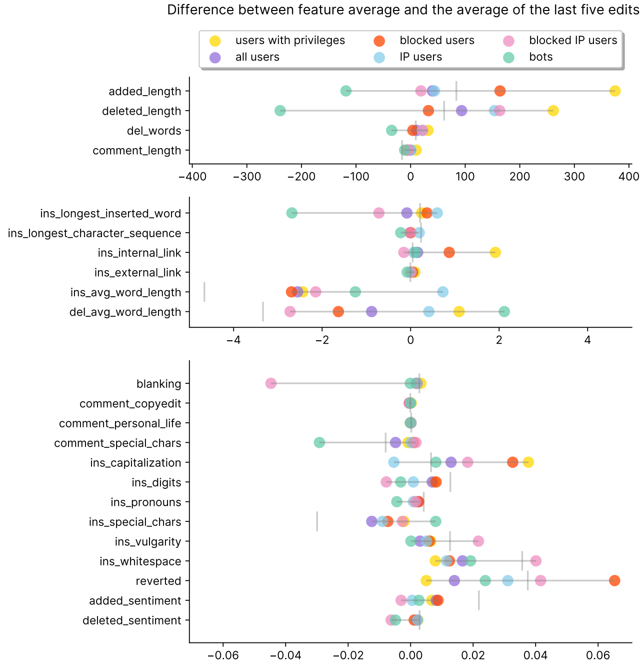


Figure 11: The difference in feature values between users' last five edits for each group and their average for all edits. The blocked users last five edits do not differ from their averages enough to make them an outlier.

We see a decrease in IP editing with a corresponding increase in user editing, suggesting that more editors are choosing to edit using an account. This is hopeful, as the quality of user edits are likely to be higher quality than those of IP editors.

Blocked users seem to edit similarly to users, which suggests they are not often blocked for "spammy" editing. This means that we cannot classify these editors easily based on their editing trends (RQ3); more advanced techniques or detailed data would be required to accurately classify them. Furthermore, a large majority of blocked users do not tend to edit for long, which is positive as it would be concerning to see many long standing and productive users getting blocked.

7 FUTURE WORK

We envision three specific avenues for future exploration of Wikipedia data using our techniques: improved classification methods, incorporating additional namespaces, and including the social domain of WikiProjects.

We will investigate more complex classification techniques, such as utilizing a Multi-layer Neural Network and computing additional features over the edit history, to effectively classify users—especially blocked users. It would also be beneficial to include features which consider additional semantic content and keywords of the edit. Secondly, we will expand benchmarking of our processes to include additional namespaces, specifically main space. By using this larger dataset, we would incorporate more user edit details, but it comes at a storage cost trade-off if inserted and deleted text for each edit is included. Expanding to main space would additionally require careful consideration of calculated features before the extraction is performed due to the computation time needed. Finally, we will investigate how WikiProjects change over time and whether there are positive or negative trends in the data. The guidance that results from these projects would help create a better encyclopedia for us all.

8 CONCLUSION

We created evocative and informative visualizations of user engagement on Wikipedia talk pages through the lens of user groups. These visualizations allowed us to see the role that special users have played throughout Wikipedia history, as well as the strong and continually engaged communities that began editing in 2006. Additionally, we show that while talk page engagement peaked in 2007 and has declined since, we can see a promising trend of rising engagement from new users in the past 4 years.

As part of this process, we have detailed a scalable method of performing research, utilizing the public XML data dumps containing the full Wikipedia edit history. We created the Namespace Database tool to create a MySQL database of Wikipedia edits for any Wikipedia namespace, category, or set of pages. This tool is fully documented and published under an open license, as so to be extensible and fork-able for future development.

Finally, we attempted to use our extracted features of user edits to classify blocked users. However, while these features provided insight into how different groups of users engage with Wikipedia talk pages, they were insufficient to isolate the editing habits of blocked users. We surmise that a more complex model or more detailed dataset would be needed to successfully classify these users.

ACKNOWLEDGMENTS

Aaron Halfaker and Lane Rasberry were key to this project's success, and we are grateful for the support that they both provided.

REFERENCES

[1] Alex Woodson. 2007. Wikipedia remains go-to site for online news. https://www.reuters.com/article/us-media-wikipedia/wikipedia-remains-go-to-site-for-online-news-idUSN0819429120070708. [Accessed 5-May-2020].

[2] Daniel Bégin, Rodolphe Devillers, and Stéphane Roche. 2018. The life cycle of contributors in collaborative online communities—the case of OpenStreetMap. *International Journal of Geographical Information Science* 32, 8 (2018), 1611–1630. https://doi.org/10.1080/13658816.2018.1458312 arXiv:https://doi.org/10.1080/13658816.2018.1458312

[3] Anamika Chhabra, Rishemjit Kaur, and S. R.S. Iyengar. 2020. Dynamics of Edit War Sequences in Wikipedia. In *Proceedings of the 16th International Symposium on Open Collaboration* (Virtual conference, Spain) *(OpenSym 2020)*. Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/3412569.3412585

[4] Christine de Kock and Andreas Vlachos. 2021. I Beg to Differ: A study of constructive disagreement in online conversations. arXiv:2101.10917 [cs.CL]

[5] Brian Everitt. 1998. *The Cambridge Dictionary of Statistics.* Cambridge University Press, Cambridge, UK ; New York.

[6] Aaron Halfaker. 2017. Mediawiki-utilities/mwxml - MediaWiki. https://www.mediawiki.org/wiki/Mediawiki-utilities/mwxml [Online; accessed 8-May-2020].

[7] Aaron Halfaker, R. Stuart Geiger, Jonathan Morgan, and John Riedl. 2013. The Rise and Decline of an Open Collaboration System: How Wikipedia's reaction to sudden popularity is causing its decline. *American Behavioral Scientist* 57, 5 (May 2013), 664–688. https://doi.org/10.1177/0002764212469365

[8] Sara Javanmardi, David W McDonald, and Cristina V Lopes. 2011. Vandalism detection in Wikipedia: a high-performing, feature-rich model and its reduction through Lasso. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration.* ACM, 82–90.

[9] Carlin MacKenzie. 2020. Namespace Database - A tool to create a database of Wikipedia edits. https://www.github.com/carlinmack/NamespaceDatabase/. https://doi.org/10.5281/zenodo.3817987

[10] Martin von Gagern. 2014. GNU Wdiff. https://www.gnu.org/software/wdiff/. [Accessed 8-May-2020].

[11] Sergio Martinez-Ortuno, Deepak Menghani, and Lars Roemheld. 2014. Sentiment as a Predictor of Wikipedia Editor Activity. (2014).

[12] Paolo Massa. 2011. Social networks of wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia.* 221–230. https://www.gnuband.org/papers/social_networks_of_wikipedia/

[13] David McCandless. 2020. Wikipedia's lamest edit wars. https://informationisbeautiful.net/visualizations/wikipedia-lamest-edit-wars/

[14] Charu Rawat, Arnab Sarkar, Sameer Singh, Rafael Alvarado, and Lane Rasberry. 2019. Automatic Detection of Online Abuse and Analysis of Problematic Users in Wikipedia. In *2019 Systems and Information Engineering Design Symposium (SIEDS).* IEEE. https://doi.org/10.1109/sieds.2019.8735592

[15] Jodi Schneider, John G Breslin, and Alexandre Passant. 2010. A content analysis: How Wikipedia talk pages are used. (2010).

[16] Julian Seward. 1996. bzip2 and libbzip2. http://sourceware.org/bzip2/ [Online; accessed 20-June-2021].

[17] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. 2009. The Singularity is Not near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (Orlando, Florida) *(WikiSym '09).* Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/1641309.1641322

[18] Dario Taraborelli and Giovanni Luca Ciampaglia. 2010. Beyond Notability. Collective Deliberation on Content Inclusion in Wikipedia. In *2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop.* 122–125. https://doi.org/10.1109/SASOW.2010.26

[19] Wikipedia contributors. 2021. Wikipedia:Database download. https://en.wikipedia.org/wiki/Wikipedia:Database_download [Online; accessed 13-June-2021].

[20] Andy B. Yoo, Morris A. Jette, and Mark Grondona. 2003. SLURM: Simple Linux Utility for Resource Management. In *Job Scheduling Strategies for Parallel Processing*, Dror Feitelson, Larry Rudolph, and Uwe Schwiegelshohn (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 44–60.

[21] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, Melbourne, Australia, 1350–1361. https://doi.org/10.18653/v1/P18-1125

A APPENDIX

A.1 Composition of special users

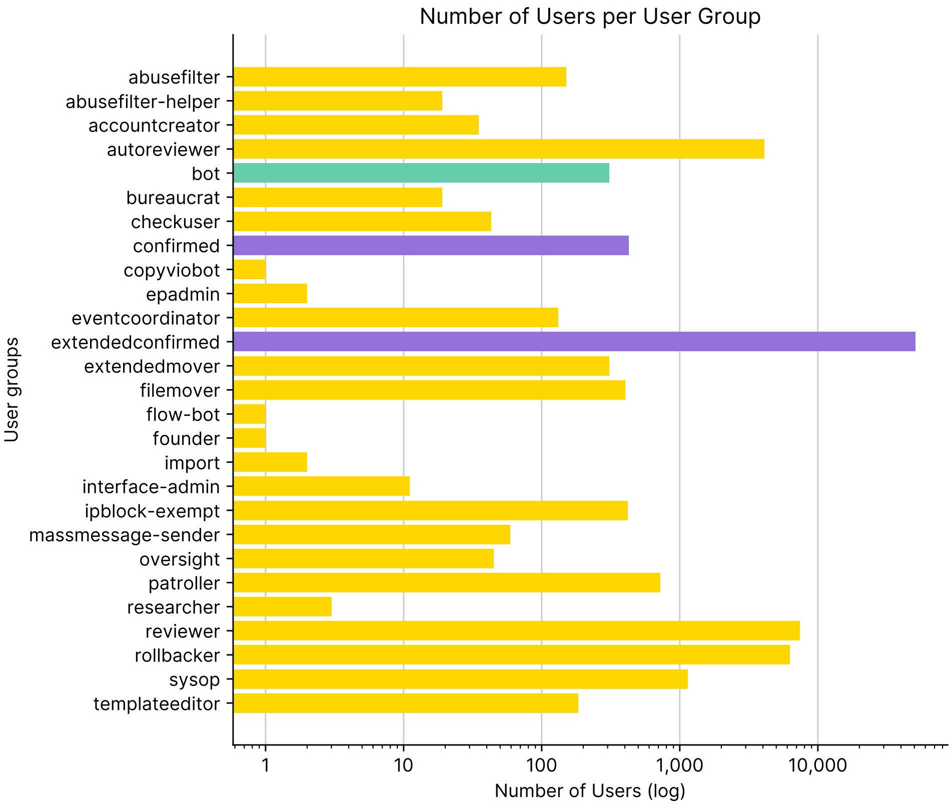


Figure 12: A plot of the number of users which have each privilege. Note the log scale when interpreting this graph. The color of the bar represents which of our groups that user belongs to.

Figure 12 illustrates the number of users with certain privileges on Wikipedia. This is the only plot which does not source from our database, but instead from the user group assignments dump[5]. All groups are positive and denote a privilege that the user can subsequently do. Details of the groups can be found on the “User access levels” page[6].

[5] https://dumps.wikimedia.org/

[6] https://en.wikipedia.org/wiki/Wikipedia:User_access_levels

WDProp: Web Application to Analyse Multilingual Aspects of Wikidata Properties

John Samuel
john.samuel@cpe.fr
CPE Lyon, LIRIS, UMR-CNRS 5205, Université de Lyon
Lyon, France

ABSTRACT

Compared to Wikipedia, Wikidata is a single domain website with the possibility to view information in multiple languages. Translation plays a significant role in Wikidata. Unlike Wikidata items, Wikidata properties are influenced less by translation bots and require a meaningful amount of human effort. The study of Wikidata property creation and translation is, therefore, very essential. Since the inception of Wikipedia, several research works have focused on the information flow among different language Wikipedias. The attention has now shifted to the way information on Wikidata is created and translated. The focus of this article is the Wikidata properties. WDProp is a web application created to understand and obtain an integrated view on the various multilingual aspects of Wikidata properties, from their proposition to their use on multiple domains.

CCS CONCEPTS

• Software and its engineering → Software libraries and repositories; **• Information systems → Collaborative and social computing systems and tools**; **Open source software**; **Data mining**.

KEYWORDS

Wikidata, Wikidata Properties, Collaboration, Multilingual Development

ACM Reference Format:
John Samuel. 2021. WDProp: Web Application to Analyse Multilingual Aspects of Wikidata Properties. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3479986.3479996

1 INTRODUCTION

Wikidata [17], a Wikimedia project created in 2012 is a free, linked, open, collaborative, and multilingual knowledge base. There are significant differences between Wikidata and Wikipedia. Unlike Wikipedia, which contains the majority of information in unstructured form (e.g., textual description), Wikidata is a structured knowledge base. Wikipedia uses multiple subdomains for different languages, each of which is managed by the associated language community. Wikidata, on the other hand, is a single domain website. All the information related to a single topic in multiple languages is referred to by a single URL. Logged-in users may be able to view the information in their local language by configuring the local language in their settings. Users can also make use of $uselang = lc$, in the URL to view the information in any other language lc, where lc is the language code. This is indeed a major change for Wikipedia users used to multiple websites for looking up information in different languages. Such an approach of building a central store for storing facts (or statements) related to multiple domains is challenging. This approach of collaborative multilingual and multi-domain ontology development on Wikidata has been the focus of many recent works [6, 8, 11].

Nevertheless, it is important to state that not all communication on Wikidata exists in a multilingual way. There exist specific pages where contributors need to use monolingual text for expressing their opinions, like in discussion and User:Talk pages. Whether a truly multilingual experience can be obtained on Wikidata is an open and relevant research question. This article focus on a small, but key aspect of Wikidata: properties to comprehend the multilingual aspects around them. Wikidata properties play a major role in describing knowledge across domains. Unlike Wikidata items, they cannot be simply created but requires discussion and voting. Depending on the need, the properties are proposed by the Wikidata contributors. Such propositions are discussed by the community members, where some may point to the availability of existing properties and others agreeing to the need for the creation of the property. Every property proposition undergoes voting, and if a property has achieved a sufficient number of supporting votes, the property may be created and made available for use.

A newly created property may not be available in all the languages and may have one or more translations based on the initial translations proposed by the property proposer. This makes it difficult for users who do not speak or comprehend any of the available translations. The role played by bilingual or multilingual speakers is therefore very important. It is also difficult to use the properties, immediately after their creation, if their usage is not properly described and documented. For this purpose, Wikidata WikiProjects for different domains have been created, which curates the properties and documents the usage for describing Wikidata items belonging to the given domain. The problem of finding relevant properties has also garnered some attention in the research community. There have been several works on predicting properties

OpenSym 2021, September 15–17, 2021, Online, Spain
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479996

[3, 18] for describing domain knowledge, including a recent work on their prediction from Wikipedia (e.g., Wiki2Prop [10]). It is important to state here that such property suggestions or predictions may require some validation by the contributors and hence their translations cannot be overlooked.

Translation of Wikidata properties is an important topic, and its analysis in real-time is equally important, given the highly evolving, open, and collaborative nature of Wikidata. Wikidata adds support for new languages and datatypes from time to time. There is no easy way to analyze the multilingual translation of Wikidata properties. There are some static pages/tools to track the datatypes and the properties, but they are not linked to languages and the property translations. In this article, we discuss the development of WDProp as a solution to these problems. WDProp bridges this gap between properties and their multilingual translations, giving the users multiple possible ways to navigate and analyze properties and their translations. WDProp also considers WikiProjects for contributors who are interested in a subset of curated information and the associated translation statistics. Section 2 gives a brief introduction of Wikidata, particularly focusing on properties. Some works on Wikidata properties are presented in section 3. Section 4 presents the development of WDProp and the associated results. Starting from some coarser analysis, it shows the results related to some fine-granular analysis. Finally, section 5 concludes the article and presents the future course of action.

2 WIKIDATA PROPERTIES

Wikidata item pages are used to describe different classes like museum, person, archive, website, conference, scientific event[1] etc. and their instances like OpenSym 2019[2], Wikidata[3] etc. In addition to Wikidata item pages, there are property pages (e.g., instance of[4], country (P17)[5], official website (P856)[6], etc.), entity schema pages (Wikimedia project [7]) and the recently introduced lexicographical pages (e.g., lexeme (L315)[8]). Properties are used to describe items (e.g, the country of OpenSym 2019 was described using P17). Entity schemas are used to validate the Wikidata properties and items. It is worth mentioning here that all these pages have an identifier of the form [QPLE][0-9][0-9]*. Unlike Wikipedia, the contributors may not know the topic of the page from the URL, without looking at the page content. For every such page, there is a discussion page (e.g., Wikidata (Q2013) talk page [9]), where contributors discuss the ways to improve the associated page.

It is difficult to remember all the identifiers of properties, items, etc., especially for usage by humans. Properties, items, and entity schemas need to be translated, i.e., these pages must have labels, descriptions, and aliases in any of the supported Wikidata languages. A property, item, and entity schema can have one label, one description, and multiple aliases in any supported language. Such translations are useful for finding the relevant items through the search interface. Thus contributors can find the relevant identifier by searching their local names.

Properties have datatypes. For example, to specify the official website of OpenSym, we may need a property with the datatype URL. Property P856 (official website) is one such example belonging to the datatype URL. There are several other datatypes like external identifiers, media, geographical coordinates, etc. Properties are not just limited to Wikidata, there are properties in other external data stores like DBPedia, VIAF, OCLC, ISBN, etc. Properties could be matched to properties of other external data stores through a special property called equivalent property[10]. Items can also be linked to entities in other external data stores through the properties belonging to the datatype: external identifier.

2.1 Property proposal and creation

Unlike Wikidata items, which require that any item created to follow the notability guidelines[11], the creation of property items is a much longer process. Figure 1 shows the process of Wikidata property creation and its possible deletion.

A contributor looking to describe an item may not find a relevant property to describe an item. In this case, they may propose a new property on the property proposal page[12]. A selected number of topics have been created for allowing the contributors to discuss and find contributors belonging to a particular domain. Some examples include authority control, person, place, transportation, etc.

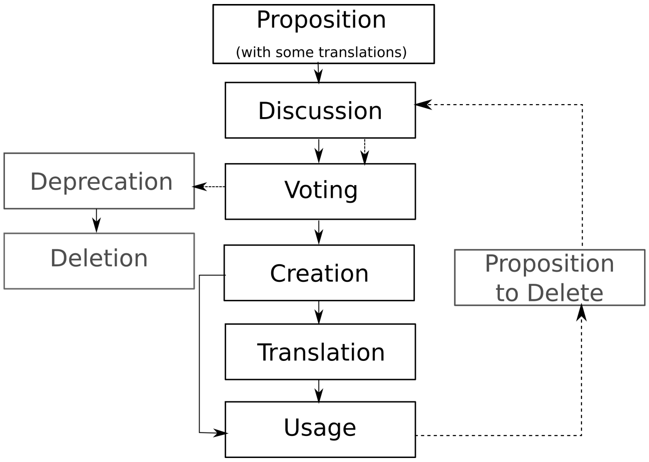


Figure 1: Property proposition, creation and possible deletion. Dark boxes and arrows shows the steps related to property creation, translation and usage. Light gray boxes and dotted arrows shows the steps related to property deletion. Some steps are shared by property creation and deletion, like discussion and voting.

An example property discussion is highlighted in Figure 2. This is the property proposal discussion of P3966 (programming paradigm)[13]. Figure 2a shows the information proposed by the property

[1] https://www.wikidata.org/wiki/Q52260246
[2] https://www.wikidata.org/wiki/Q56259215
[3] https://www.wikidata.org/wiki/Q2013
[4] https://www.wikidata.org/wiki/Property:P31
[5] https://www.wikidata.org/wiki/Property:P17
[6] https://www.wikidata.org/wiki/Property:P856
[7] https://www.wikidata.org/wiki/EntitySchema:E2
[8] https://www.wikidata.org/wiki/Lexeme:L315
[9] https://www.wikidata.org/wiki/Talk:Q2013
[10] https://www.wikidata.org/wiki/Property:P1628
[11] https://www.wikidata.org/wiki/Wikidata:Notability
[12] https://www.wikidata.org/wiki/Wikidata:Property_proposal
[13] https://www.wikidata.org/wiki/Wikidata:Property_proposal/programming_paradigm

proposer. Figure 2b shows the voting and in Figure 2b, we see the source code of the property proposition.

Readers may refer to the French[14] or the Spanish[15] rendering of this property discussion page. We see that some of the information like the French or the Spanish translations are available for some information like Support, Oppose, Comment. These template pages support,oppose, comment have already been translated in these languages.

As can be seen, the property proposer (or/and some other contributors) translated the property label and description in some languages. For this purpose, TranslateThis template has been used. At the time of property creation, these labels and descriptions made available, as seen in Figure 1. Properties may be translated into other languages after creation. It should be stressed that the labels, descriptions, and aliases may be modified during the lifespan of a property. Some of these modifications may be possible vandalism and require to be detected[16]. Finally, the lifespan of a property may be limited. Sometimes, contributors may propose that a property should be deleted[16]. This may be because of the need to change the datatype or because a new and better property has been proposed with which the concerned property can be merged.

2.2 Property curation: WikiProjects and Property Classes

Another major problem is the difficulty in finding relevant property items for describing an item. Take for example, the following three properties:

- country (P17) used for specifying the country of the item.
- country of origin (P495) used for specifying the country of origin of the item, when the item is a creative work, food, etc.
- country of citizenship (P27) for specifying the nationality of a person (item).

A new contributor may not be aware of all such properties, and often may end up using P17 (country) for specifying the country of origin of a creative work. For helping the contributors, there are Wikidata WikiProjects[17] that curate the properties relevant to a particular domain. For example, there are WikiProjects related to books[18], museums[19], etc. Another possible way is to curate them using the property: properties for this type (P1963), that can be used to find the properties for a given type (a class etc.)

2.3 Property translation and evolution

As seen in Figure 1, Wikidata properties constantly evolve. Properties are created, modified, translated and even deleted. New languages, datatypes, and WikiProjects are also created from time to time. To improve the language coverage and coverage of topics from across the world, the translation of properties cannot be overlooked. Property translation and evolution is the focus of this article.

3 STATE OF THE ART

Wikidata provides a SPARQL endpoint and Mediawiki API for accessing the structured data as well as the commit history. A number of tools have been created during the past few years[20], which help contributors to describe, analyse and visualize Wikidata items. Tools like PropBrowser[21], SQID property browser[22] focus on Wikidata properties. Some commonly used tools on Wikidata by contributors for generating lists are Listeria[23] and InteGraality[24]. This has been used by certain contributors to maintain a list of datatypes[25]. However, these tools generate the lists in a periodic manner (like weekly updates etc.)

Collaborative [11] and multilingual aspects [6] of Wikidata has been the focus of many recent works. Many of these works make use of periodic Wikidata datadumps, focusing on the global coarser analysis. But the contributors focus on some properties or a subset of curated properties for their works. Certain visualizations like the visualization of deletion discussions considered by Notabilia [15] or information flow across different language Wikipedias [1, 4] give interesting insights. While analysing and visualizing the translations of Wikidata, one may observe the multilingual nature of the contributors and also the role played by the bots[5]. Some reccent work have focuses on the stability of property labels[16]. It must be noted that the property labels are meant for human consumption and undesired or possible vandalism on the properties may produce unexpected semantics of information. Hence it is important to obtain real-time and fine-granular analysis of Wikidata properties.

Translation of properties may aid to extracting structured information from Wikipedia articles to populate infoboxes [9] and Wikidata[26] [27]. Finally, it is equally important to point the role played by Wikidata WikiProjects [7, 8] in curation of properties. The focus of WDProp proposed in this article is on Wikidata properties and the multilingual aspects around the properties. The goal is to navigate and explore properties by languages, datatypes, WikiProjects, etc.

4 WDPROP: FROM COARSER ANALYSIS TO FINE-TUNED ANALYSIS

One approach for analysis of Wikidata items and properties is to download Wikidata datadump[28] with the commit history for the study. Another approach is to make use of SPARQL endpoints and Wikidata Mediawiki API, especially when the focus is on a small subset of information. *Multilingual Wikidata Property Translation Flow Dataset* [2] was created using the latter approach. Datadumps, complete or selective are quite useful for coarser analysis and may provide some useful insights [6, 11].

Table 1 shows the details of *Multilingual Wikidata Property Translation Flow Dataset*. This dataset contains the translation flow of

[14] https://www.wikidata.org/wiki/Wikidata:Property_proposal/programming_paradigm?uselang=fr

[15] https://www.wikidata.org/wiki/Wikidata:Property_proposal/programming_paradigm?uselang=es

[16] https://www.wikidata.org/wiki/Wikidata:Properties_for_deletion

[17] https://www.wikidata.org/wiki/Wikidata:WikiProject

[18] https://www.wikidata.org/wiki/Wikidata:WikiProject_Books

[19] https://www.wikidata.org/wiki/Wikidata:WikiProject_Museums

[20] https://hay.toolforge.org/directory/

[21] http://tools.wmflabs.org/hay/propbrowse

[22] https://sqid.toolforge.org/#/browse?type=properties

[23] https://www.wikidata.org/wiki/Wikidata:Listeria

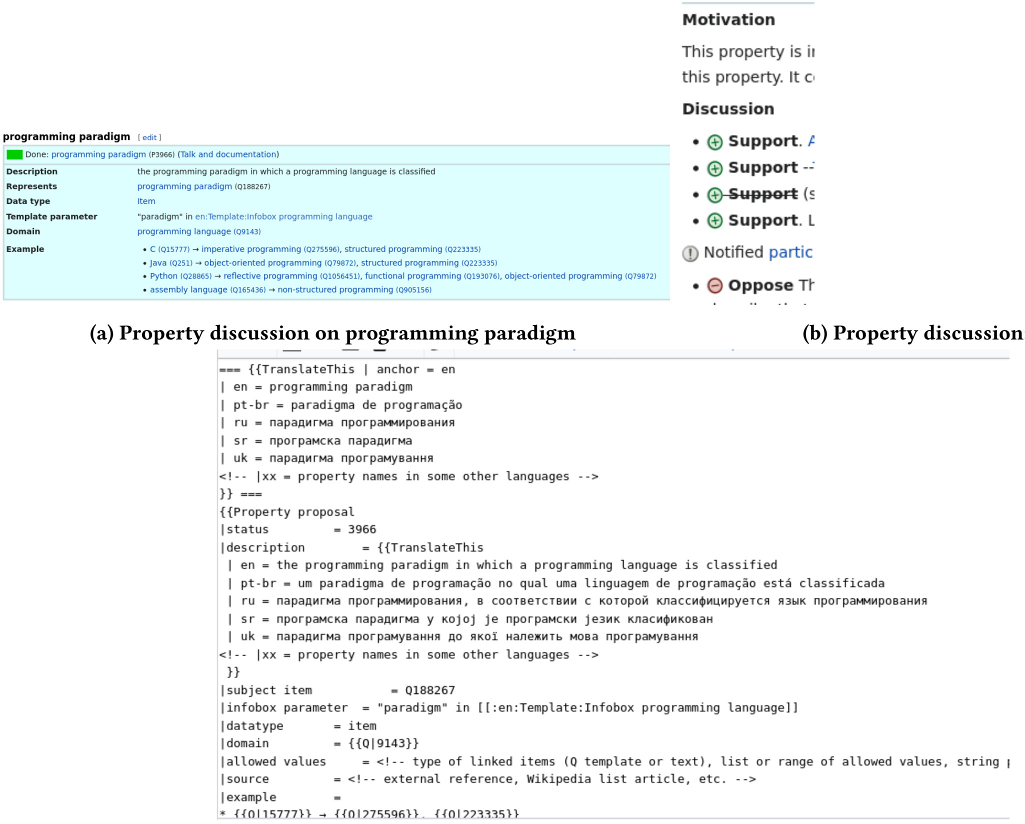
[24] https://integraality.toolforge.org/

[25] https://www.wikidata.org/wiki/Wikidata:WikiProject_Properties/Reports/Datatypes

[26] https://pltools.toolforge.org/harvesttemplates/

[27] https://petscan.wmflabs.org/

[28] https://www.wikidata.org/wiki/Wikidata:Database_download



programming paradigm [edit]

Done: programming paradigm (P3966) (Talk and documentation)

Description	the programming paradigm in which a programming language is classified
Represents	programming paradigm (Q188267)
Data type	Item
Template parameter	"paradigm" in en:Template:Infobox programming language
Domain	programming language (Q9143)
Example	• C (Q15777) → imperative programming (Q275596), structured programming (Q223335) • Java (Q251) → object-oriented programming (Q79872), structured programming (Q223335) • Python (Q28865) → reflective programming (Q1056451), functional programming (Q193076), object-oriented programming (Q79872) • assembly language (Q165436) → non-structured programming (Q905156)

(a) Property discussion on programming paradigm

Motivation

This property is i
this property. It c

Discussion

- ⊕ **Support**. A
- ⊕ **Support** --
- ⊕ ~~**Support**~~ (s
- ⊕ **Support**. L

ⓘ Notified partic

- ⊖ **Oppose** Th

(b) Property discussion: Votes

```
=== {{TranslateThis | anchor = en
| en = programming paradigm
| pt-br = paradigma de programação
| ru = парадигма программирования
| sr = програмска парадигма
| uk = парадигма програмування
<!-- |xx = property names in some other languages -->
}} ===
{{Property proposal
|status         = 3966
|description        = {{TranslateThis
 | en = the programming paradigm in which a programming language is classified
 | pt-br = um paradigma de programação no qual uma linguagem de programação está classificada
 | ru = парадигма программирования, в соответствии с которой классифицируется язык программирования
 | sr = програмска парадигма у којој је програмски језик класификован
 | uk = парадигма програмування до якої належить мова програмування
<!-- |xx = property names in some other languages -->
 }}
|subject item           = Q188267
|infobox parameter  = "paradigm" in [[:en:Template:Infobox programming language]]
|datatype       = item
|domain         = {{Q|9143}}
|allowed values     = <!-- type of linked items (Q template or text), list or range of allowed values, string p
|source         = <!-- external reference, Wikipedia list article, etc. -->
|example        =
* {{Q|15777}} → {{Q|275596}}, {{Q|223335}}
```

(c) Property discussion template

Figure 2: Property proposition and discussion of P3966 (programming paradigm)

labels, descriptions, and aliases of Wikidata properties collected on July 7, 2019. The dataset consists of four columns: timestamp of the translation, the identifier of the property translated, the language of translation, and the type of translation (label, description, or alias). The 6347 properties considered in the study showed that a mean of 21.44 property labels in different languages was available. The minimum and the maximum number of property labels are 1 and 154 respectively. The distribution of the count of properties and the number of languages is shown in Figure 3.

Measure	Value
count	6347
mean	21.444935
std	20.897675
min	1.000000
max	154.000000

Table 1: Details of *Multilingual Wikidata Property Translation Flow Dataset* [2]. The dataset contains the details of every type of translation (labels, description and aliases) on 6347 properties.

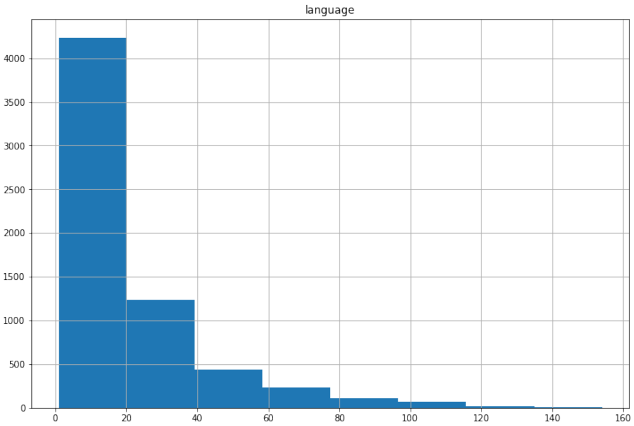


Figure 3: Distribution of labels in different languages. The X-axis corresponds to the number of languages and the Y-axis corresponds to the number of properties. Very less number of properties have label translations in more than 100 languages.

Further analysis of these property labels was made on the count of combined occurrences of languages on different properties and the results of the top seven combinations are shown in Table 2.

Count	Language Combinations	Count	Language Combinations	Count	Language Combinations
6347	en	6259	ar, uk	6257	en, nl, ar, uk
6338	ar	5923	fr, ar	5919	en, nl, fr, ar
6306	nl	5905	fr, uk	5903	en, nl, fr, uk
6263	uk	4819	ar, ca	5901	en, ar, fr, uk
5928	fr	4795	ca, uk	5899	nl, fr, uk, ar
4823	ca	4530	fr, ca	4818	en, nl, ar, ca
4019	de	4015	ar, de	4794	en, nl, ca, uk

Table 2: Total number of properties with the label translations in the language combinations (Top 7)

For example, considering a combination of four languages, English (en), Dutch (nl), Arabic (ar) and Ukrainian (uk) property labels are available on 6257 properties.

The above dataset can also be used to study the property translation flow for a given property as shown in Figure 4. It took a period between 2013 and 2019 for the translation of the labels (violet dots) of the property P17 (country) to be available in multiple languages. The first descriptions (light blue dots) were subsequently translated and the first aliases (red dots) were added after a long time. Check the distances between the violet, blue and red dots for some languages. By focusing on modifications (i.e, changes made after the first translations), a modified form of this visualization can be used to detect possible vandalisms.

However, the above analysis is limited to a snapshot of data at a given time. Considering the rapidly evolving nature of Wikidata, there is a need to have some real-time view of Wikidata properties. WDProp was developed to understand and obtain an integrated view on the different aspects of properties, their creation, and translation. The focus here is on obtaining real-time statistics, on a much fine-granular basis, i.e., instead of focusing on the global analysis of properties (as seen above), the users can focus on a single property or a combination of curated properties. WDProp provides a visual interface to such information.

Property translations can be analyzed in different ways. There are two major translations: property labels and property descriptions. Properties may also have some aliases in certain languages. Whether all the aliases in one language need to be available in other languages is an open question. For a given language, the language community may wish to ensure the translation of all the property labels and descriptions. For some generic properties like country (P17), the goal of the Wikidata community members is to ensure the translation in all the supported languages.

WDProp separates property translations in the following manner:

- translated and untranslated property labels
- translated and untranslated property descriptions
- translated and untranslated property aliases

A Wikidata contributor may find one or more of the above pieces of information relevant, some of whom may wish to obtain the latest translation statistics, and some others looking for properties not yet translated.

4.1 Development

WDProp was developed using basic web technologies like HTML, Javascript, and CSS. It makes use of the SPARQL queries and Wikidata MediaWiki API to obtain the latest information from the Wikidata servers. It, therefore, does not require any installation since it contains a collection of HTML, Javascript, and CSS files. This simple approach ensures that it can be tested on any modern browser, including desktop and mobile devices. It is also available on Toolforge[29], for users wishing to use links that can be shared with others.

4.2 Results

WDProp can be used to analyze the collaborative approach to the development of multilingual ontology [12] and to obtain the real-time information related to multilingual aspects of Wikidata properties, especially their translation [13]. It provides the following information and features:

- **Bookmarkable links**: This feature ensures that the links are bookmarkable and use Wikidata supported codes and identifiers for obtaining the information related to languages, properties, datataypes, etc.
- **List of supported languages**: see Figure 5
- **Translation statistics of labels, descriptions and aliases of Wikidata properties**: see Figure 6
- **List of properties**: see Figure 8
- **Compare translation statistics among different languages**: see Figure 7
- **Use of references and equivalent properties**: see Figure 9.
- **Translation statistics in a given language**: see Figure 10 and Figure 11.
- **Translation statistics of property discussion templates**: see Figure 12
- **Navigation of properties by datatypes**: see Figure 13
- **Navigation of properties by classes**: see Figure 14
- **Navigation of properties by Wikidata WikiProjects**: see Figure 17
- **Visualize path of translation**: see Figure 18

In this article, the URLs used for the above screenshots are given as footnotes.

Figure 5 shows the list of supported languages[30] on Wikidata. New languages are incubated on Wikidata. Hence the number of

[29] https://wdprop.toolforge.org/

[30] https://wdprop.toolforge.org/languages.html

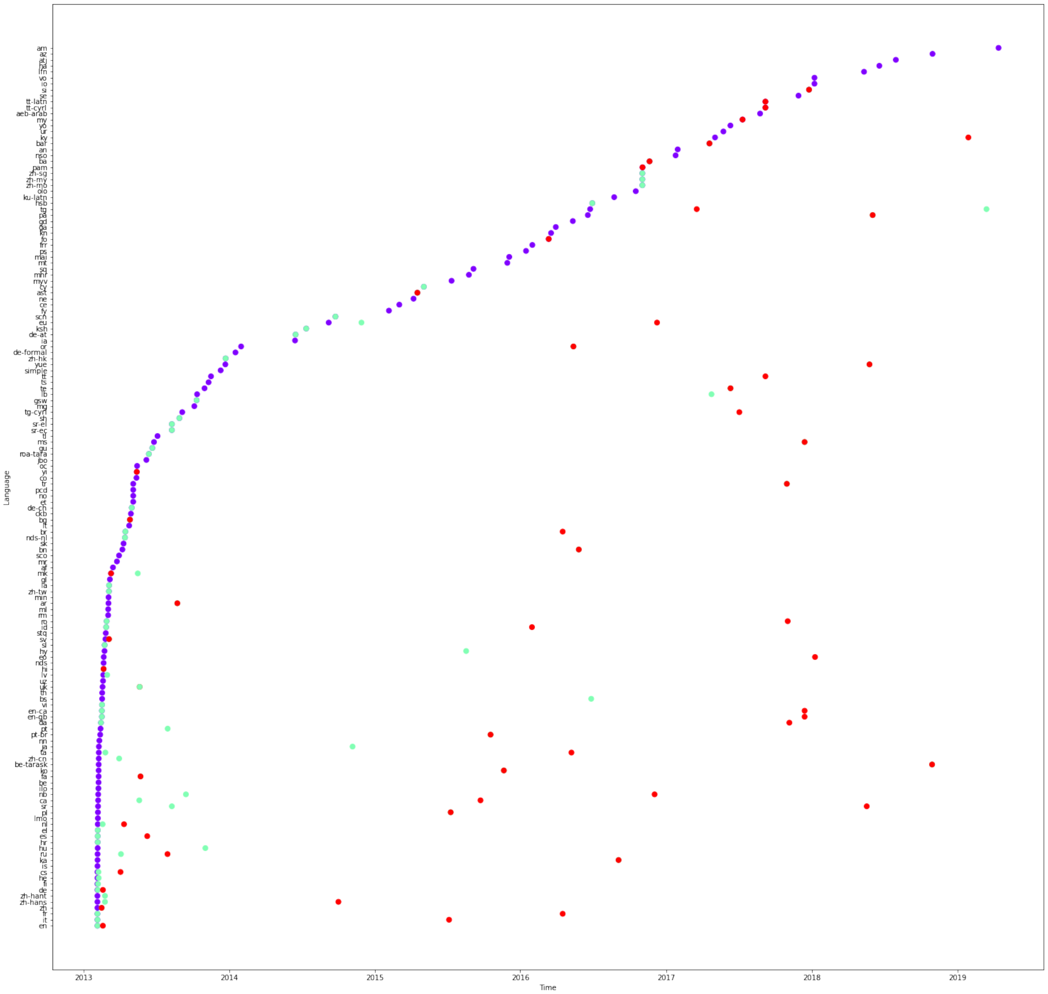


Figure 4: Timeline of translation of labels (violet dots), descriptions (blue dots) and aliases (red dots) of property P17 (country). The Y-axis shows the languages and the Y-axis shows the period between the year 2013 and 2019.

supported languages on Wikidata has evolved over the years. This list is useful to know the number of supported languages at any given time.

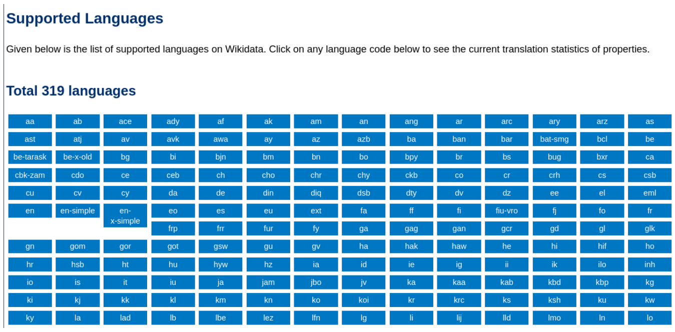
Property translation statistics[31] is given in Figure 6. It shows the number of labels, descriptions, and aliases in each of the supporting languages seen above. It is also possible to obtain the number of labels, descriptions, and aliases missing translation[32].

Comparison of property translation statistics is given in Figure 7. It shows the property translation statistics English(en), French (fr) and Spanish (es)[33]. The users can choose a selection of languages

[31] https://wdprop.toolforge.org/translated.html

[32] https://wdprop.toolforge.org/untranslated.html

[33] https://wdprop.toolforge.org/compare.html?languages=en,%20fr,%20es



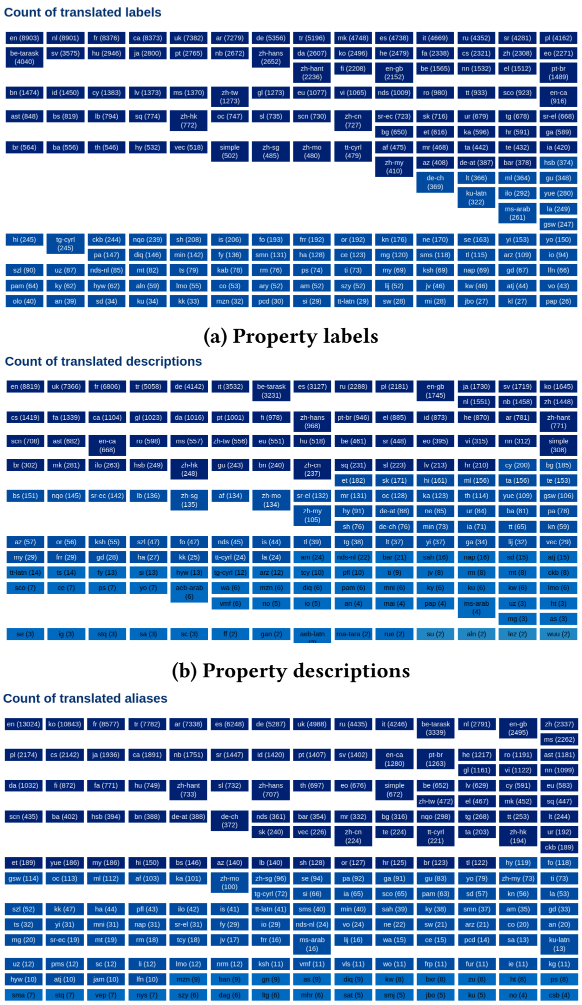
Supported Languages

Given below is the list of supported languages on Wikidata. Click on any language code below to see the current translation statistics of properties.

Total 319 languages

aa	ab	ace	ady	af	ak	am	an	ang	ar	arc	ary	arz	as
ast	atj	av	avk	awa	ay	az	azb	ba	ban	bar	bat-smg	bcl	be
be-tarask	be-x-old	bg	bi	bjn	bm	bn	bo	bpy	br	bs	bug	bxr	ca
cbk-zam	cdo	ce	ceb	ch	cho	chr	chy	ckb	co	cr	crh	cs	csb
cu	cv	cy	da	de	din	diq	dsb	dty	dv	dz	ee	el	eml
en	en-simple	en-x-simple	eo	es	eu	ext	fa	ff	fi	fiu-vro	fj	fo	fr
			frp	frr	fur	fy	ga	gag	gan	gcr	gd	gl	glk
gn	gom	gor	got	gsw	gu	gv	ha	hak	haw	he	hi	hif	ho
hr	hsb	ht	hu	hyw	hz	ia	id	ie	ig	ii	ik	ilo	inh
io	is	it	iu	ja	jam	jbo	jv	ka	kaa	kab	kbd	kbp	kg
ki	kj	kk	kl	km	kn	ko	koi	kr	krc	ks	ksh	ku	kw
ky	la	lad	lb	lbe	lez	lfn	lg	li	lij	lld	lmo	ln	lo

Figure 5: Wikidata supported languages



Count of translated labels

en (8903) nl (8901) fr (8376) ca (8373) uk (7382) ar (7279) de (5356) tr (5196) mk (4748) es (4738) it (4669) ru (4352) sr (4281) pl (4162)
be-tarask (4040) sv (3575) hu (2946) ja (2800) pt (2765) nb (2672) zh-hans (2652) da (2607) ko (2496) he (2479) fa (2338) cs (2321) zh (2308) eo (2271)
zh-hant (2236) fi (2208) en-gb (2152) be (1565) nn (1532) el (1512) pt-br (1489)
bn (1474) id (1450) cy (1383) lv (1373) ms (1370) zh-tw (1273) gl (1273) eu (1077) vi (1065) nds (1009) ro (980) tt (933) sco (923) en-ca (916)
ast (848) bs (819) lb (794) sq (774) zh-hk (772) oc (747) sl (735) scn (730) zh-cn (727) sr-ec (723) sk (716) ur (679) tg (678) sr-el (668)
bg (650) et (616) ka (596) hr (591) ga (589)
br (564) ba (556) th (546) hy (532) vec (518) simple (502) zh-sg (485) zh-mo (480) tt-cyrl (479) af (475) mr (468) ta (442) te (432) ia (420)
zh-my (410) az (408) de-at (387) bar (378) hsb (374)
de-ch (369) lt (366) ml (364) gu (348)
ku-latn (322) ilo (292) yue (280)
ms-arab (261) la (249)
gsw (247)
hi (245) tg-cyrl (245) ckb (244) nqo (239) sh (208) is (206) fo (193) frr (192) or (192) kn (176) ne (170) se (163) yi (153) yo (150)
pa (147) diq (146) min (142) fy (136) smn (131) ha (128) ce (123) mg (120) sms (118) tl (115) arz (109) io (94)
szl (90) uz (87) nds-nl (85) mt (82) ts (79) kab (78) rm (76) ps (74) ti (73) my (69) ksh (69) nap (69) gd (67) lfn (66)
pam (64) ky (62) hyw (62) aln (59) lmo (55) co (53) ary (52) am (52) szy (52) lij (52) jv (46) kw (46) atj (44) vo (43)
olo (40) an (39) sd (34) ku (34) kk (33) mzn (32) pcd (30) si (29) tt-latn (29) sw (28) mi (28) jbo (27) kl (27) pap (26)

(a) Property labels

Count of translated descriptions

en (8819) uk (7366) fr (6806) tr (5058) de (4142) it (3532) be-tarask (3231) es (3127) ru (2288) pl (2181) en-gb (1745) ja (1730) sv (1719) ko (1645)
nl (1551) nb (1458) zh (1448)
cs (1419) fa (1339) ca (1104) gl (1023) da (1016) pt (1001) fi (978) zh-hans (968) pt-br (946) el (885) id (873) he (870) ar (781) zh-hant (771)
scn (708) ast (682) en-ca (668) ro (598) ms (557) zh-tw (556) eu (551) hu (518) be (461) sr (448) eo (395) vi (315) nn (312) simple (308)
br (302) mk (281) ilo (263) hsb (249) zh-hk (248) gu (243) bn (240) zh-cn (237) sq (231) sl (223) lv (213) hr (210) cy (200) bg (185)
et (182) sk (171) hi (161) ml (156) ta (156) te (153)
bs (151) nqo (145) sr-ec (142) lb (136) zh-sg (135) af (134) zh-mo (134) sr-el (132) mr (131) oc (128) ka (123) th (114) yue (109) gsw (106)
zh-my (105) hy (91) de-at (88) ne (85) ur (84) ba (81) pa (78)
sh (76) de-ch (76) min (73) ia (71) tt (65) kn (59)
az (57) or (56) ksh (55) szl (47) fo (47) nds (45) is (44) tl (39) tg (38) lt (37) yi (37) ga (34) lij (32) vec (29)
my (29) frr (29) gd (28) ha (27) kk (25) tt-cyrl (24) la (24) am (24) nds-nl (22) bar (21) sah (16) nap (16) sd (15) atj (15)
tt-latn (14) ts (14) fy (13) si (13) hyw (13) tg-cyrl (12) arz (12) tcy (10) pfl (10) ti (9) jv (8) rm (8) mt (8) ckb (8)
sco (7) ce (7) ps (7) yo (7) aeb-arab (6) wa (6) mzn (6) diq (6) pam (6) mni (6) ky (6) ku (6) kw (6) lmo (6)
vmf (6) no (5) io (5) an (4) mai (4) pap (4) ms-arab (4) uz (3) ht (3)
mg (3) as (3)
se (3) ig (3) stq (3) sa (3) sc (3) ff (2) gan (2) aeb-latn roa-tara (2) rue (2) su (2) aln (2) lez (2) wuu (2)

(b) Property descriptions

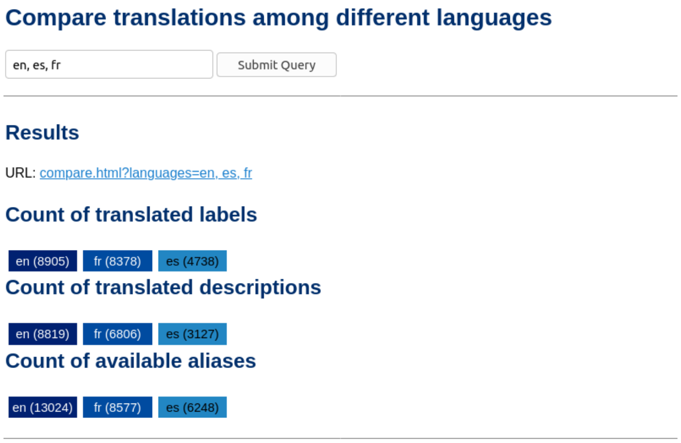
Count of translated aliases

en (13024) ko (10843) fr (8577) tr (7782) ar (7338) es (6248) de (5287) uk (4988) ru (4435) it (4246) be-tarask (3339) nl (2791) en-gb (2495) zh (2337)
ms (2262)
pl (2174) cs (2142) ja (1936) ca (1891) nb (1751) sr (1447) id (1420) pt (1407) sv (1402) en-ca (1280) pt-br (1263) he (1217) ro (1191) ast (1181)
gl (1161) vi (1122) nn (1099)
da (1032) fi (872) fa (771) hu (749) zh-hant (733) sl (732) zh-hans (707) th (697) eo (676) simple (672) be (652) lv (629) cy (591) eu (583)
zh-tw (472) el (467) mk (452) sq (447)
scn (435) ba (402) hsb (394) bn (388) de-at (388) de-ch (372) nds (361) bar (354) mr (332) bg (316) nqo (298) tg (268) tt (253) lt (244)
sk (240) vec (226) zh-cn (224) te (224) tt-cyrl (221) ta (203) zh-hk (194) ur (192)
ckb (189)
et (189) yue (186) my (186) hi (150) bs (146) az (140) lb (140) sh (128) or (127) hr (125) br (123) tl (122) hy (119) fo (118)
gsw (114) oc (113) ml (112) af (103) ka (101) zh-mo (100) zh-sg (96) se (94) pa (92) ga (91) gu (83) yo (79) zh-my (73) ti (73)
tg-cyrl (72) si (66) ia (65) sco (65) pam (63) sd (57) kn (56) la (53)
szl (52) kk (47) ha (44) pfl (43) ilo (42) is (41) tt-latn (41) sms (40) min (40) sah (39) ky (38) smn (37) am (35) gd (33)
ts (32) yi (31) mni (31) nap (31) sr-el (31) fy (29) io (29) nds-nl (24) vo (24) ne (22) sw (21) arz (21) co (20) an (20)
mg (20) sr-ec (19) mt (19) rm (18) tcy (18) jv (17) frr (16) ms-arab (16) lij (16) wa (15) ce (15) pcd (14) sa (13) ku-latn (13)
uz (12) pms (12) sc (12) li (12) lmo (12) nrm (12) ksh (11) vmf (11) vls (11) wo (11) frp (11) fur (11) ie (11) kg (11)
hyw (10) atj (10) jam (10) lfn (10) mzn (9) ban (9) gn (9) as (9) diq (9) kw (8) bxr (8) zu (8) ht (8) ps (8)
sma (7) stq (7) vep (7) nys (7) szy (6) dag (6) ltg (6) mhr (6) sat (5) smj (5) jbo (5) ku (5) no (4) csb (4)

(c) Property aliases

Figure 6: WDProp Translation Statistics of Properties: Count of available translations

and compare the translation statistics of labels, descriptions and aliases.



Compare translations among different languages

en, es, fr

Results

URL: compare.html?languages=en, es, fr

Count of translated labels

en (8905) fr (8378) es (4738)

Count of translated descriptions

en (8819) fr (6806) es (3127)

Count of available aliases

en (13024) fr (8577) es (6248)

Figure 7: Comparison of property translation statistics on WDProp: English(en), French (fr) and Spanish (es)

As discussed above, new properties are regularly proposed and voted by the community members. The properties that have been created have an associated identifier. In Figure 8, 100 properties are shown including the option to view the deleted properties[34]. It is interesting to observe here that many of the initial properties have been deleted.

Some properties have equivalent properties in other knowledge bases. Take, for example, Property (P17) is also available on DBpedia[35]. Wikidata tracks such equivalent properties, which can be considered as a quality metric for property relevance. Additionally, like Wikidata items, Wikidata properties may also have references to the different statements. However, not all properties have references. In Figure 9, the usage of equivalent properties and references[36] are shown.

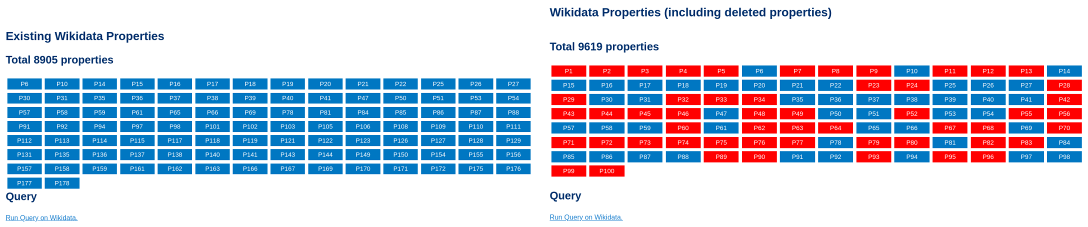
It is also possible to obtain the translation statistics for a given language. In Figure 10, current information of property translations in the Afar (aa) language[37] can be seen.

In Figure 11, current information of property translations in English (en) language[38] can be seen. When compared to Figure 10, we observe the differences in property translation in the two languages. Similar observations can be made between the English language and the languages with few speakers (or few contributors on Wikidata).

Figure 12 shows the current translation statistics of the key templates used for property discussions[39]. The numbers are lower than the total number of supported languages on Wikidata, for all four templates.

Wikidata supports several datatypes: Math, URLs, external identifiers, etc. New datatypes like wikibase:WikibaseLexeme and

[34] https://wdprop.toolforge.org/properties.html
[35] http://dbpedia.org/ontology/country
[36] https://wdprop.toolforge.org/provenance.html
[37] https://wdprop.toolforge.org/language.html?language=aa
[38] https://wdprop.toolforge.org/language.html?language=en
[39] https://wdprop.toolforge.org/templates/translated.html



Existing Wikidata Properties

Total 8905 properties

P6	P10	P14	P15	P16	P17	P18	P19	P20	P21	P22	P25	P26	P27
P30	P31	P35	P36	P37	P38	P39	P40	P41	P47	P50	P51	P53	P54
P57	P58	P59	P61	P65	P66	P69	P78	P81	P84	P85	P86	P87	P88
P91	P92	P94	P97	P98	P101	P102	P103	P105	P106	P108	P109	P110	P111
P112	P113	P114	P115	P117	P118	P119	P121	P122	P123	P126	P127	P128	P129
P131	P135	P136	P137	P138	P140	P141	P143	P144	P149	P150	P154	P155	P156
P157	P158	P159	P161	P162	P163	P166	P167	P169	P170	P171	P172	P175	P176
P177	P178												

Query

Run Query on Wikidata.

(a) Properties

Wikidata Properties (including deleted properties)

Total 9619 properties

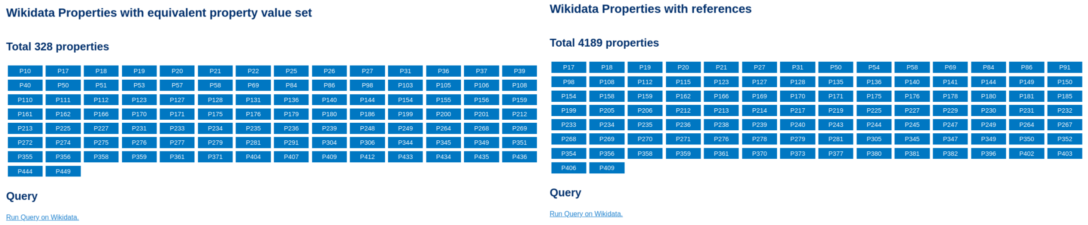
P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28
P29	P30	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	P41	P42
P43	P44	P45	P46	P47	P48	P49	P50	P51	P52	P53	P54	P55	P56
P57	P58	P59	P60	P61	P62	P63	P64	P65	P66	P67	P68	P69	P70
P71	P72	P73	P74	P75	P76	P77	P78	P79	P80	P81	P82	P83	P84
P85	P86	P87	P88	P89	P90	P91	P92	P93	P94	P95	P96	P97	P98
P99	P100												

Query

Run Query on Wikidata.

(b) Properties including deleted properties

Figure 8: List of properties on WDProp: limited to view of 100 properties



Wikidata Properties with equivalent property value set

Total 328 properties

P10	P17	P18	P19	P20	P21	P22	P25	P26	P27	P31	P36	P37	P39
P40	P50	P51	P53	P57	P58	P69	P84	P86	P98	P103	P105	P106	P108
P110	P111	P112	P123	P127	P128	P131	P136	P140	P144	P154	P155	P156	P159
P161	P162	P166	P170	P171	P175	P176	P179	P180	P186	P199	P200	P201	P212
P213	P225	P227	P231	P233	P234	P235	P236	P239	P248	P249	P264	P268	P269
P272	P274	P275	P276	P277	P279	P281	P291	P304	P306	P344	P345	P349	P351
P355	P356	P358	P359	P361	P371	P404	P407	P409	P412	P433	P434	P435	P436
P444	P449												

Query

Run Query on Wikidata.

(a) Properties with equivalent properties

Wikidata Properties with references

Total 4189 properties

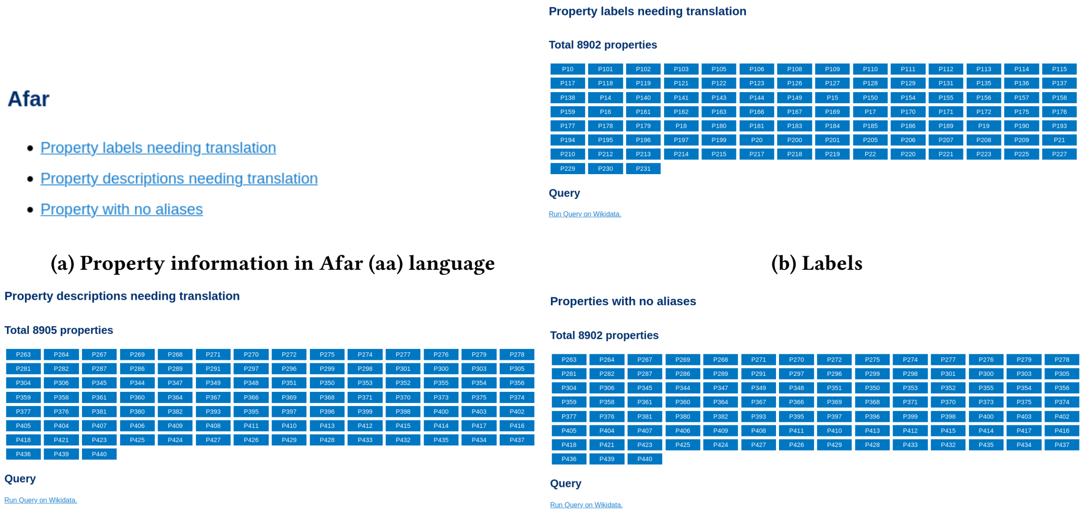
P17	P18	P19	P20	P21	P27	P31	P50	P54	P58	P69	P84	P86	P91
P98	P108	P112	P115	P123	P127	P128	P135	P136	P140	P141	P144	P149	P150
P154	P158	P159	P162	P166	P169	P170	P171	P175	P176	P178	P180	P181	P185
P199	P205	P206	P212	P213	P214	P217	P219	P225	P227	P229	P230	P231	P232
P233	P234	P235	P236	P238	P239	P240	P243	P244	P245	P247	P249	P264	P267
P268	P269	P270	P271	P276	P278	P279	P281	P305	P345	P347	P349	P350	P352
P354	P356	P358	P359	P361	P370	P373	P377	P380	P381	P382	P396	P402	P403
P406	P409												

Query

Run Query on Wikidata.

(b) Properties with references

Figure 9: Equivalent properties and usage of references on Wikidata properties



Afar

- Property labels needing translation
- Property descriptions needing translation
- Property with no aliases

(a) Property information in Afar (aa) language

Property labels needing translation

Total 8902 properties

P10	P101	P102	P103	P105	P106	P108	P109	P110	P111	P112	P113	P114	P115
P117	P118	P119	P121	P122	P123	P126	P127	P128	P129	P131	P135	P136	P137
P138	P14	P140	P141	P143	P144	P149	P15	P150	P154	P155	P156	P157	P158
P159	P16	P161	P162	P163	P166	P167	P169	P17	P170	P171	P172	P175	P176
P177	P178	P179	P18	P180	P181	P183	P184	P185	P186	P189	P19	P190	P193
P194	P195	P196	P197	P199	P20	P200	P201	P205	P206	P207	P208	P209	P21
P210	P212	P213	P214	P215	P217	P218	P219	P22	P220	P221	P223	P225	P227
P229	P230	P231											

Query

Run Query on Wikidata.

(b) Labels

Property descriptions needing translation

Total 8905 properties

P263	P264	P267	P269	P268	P271	P270	P272	P275	P274	P277	P276	P279	P278
P281	P282	P287	P286	P289	P291	P297	P296	P299	P298	P301	P300	P303	P305
P304	P306	P345	P344	P347	P349	P348	P351	P350	P353	P352	P355	P354	P356
P359	P358	P361	P360	P364	P367	P366	P369	P368	P371	P370	P373	P375	P374
P377	P376	P381	P380	P382	P393	P395	P397	P396	P399	P398	P400	P403	P402
P405	P404	P407	P406	P409	P408	P411	P410	P413	P412	P415	P414	P417	P416
P418	P421	P423	P425	P424	P427	P426	P429	P428	P433	P432	P435	P434	P437
P436	P439	P440											

Query

Run Query on Wikidata.

(c) Descriptions

Properties with no aliases

Total 8902 properties

P263	P264	P267	P269	P268	P271	P270	P272	P275	P274	P277	P276	P279	P278
P281	P282	P287	P286	P289	P291	P297	P296	P299	P298	P301	P300	P303	P305
P304	P306	P345	P344	P347	P349	P348	P351	P350	P353	P352	P355	P354	P356
P359	P358	P361	P360	P364	P367	P366	P369	P368	P371	P370	P373	P375	P374
P377	P376	P381	P380	P382	P393	P395	P397	P396	P399	P398	P400	P403	P402
P405	P404	P407	P406	P409	P408	P411	P410	P413	P412	P415	P414	P417	P416
P418	P421	P423	P425	P424	P427	P426	P429	P428	P433	P432	P435	P434	P437
P436	P439	P440											

Query

Run Query on Wikidata.

(d) Aliases

Figure 10: Property information in the Afar (aa) language

wikibase:WikibaseSense were recently created for representing lexicographical data. Figure 13 shows the list of available datatypes[40] and the properties belonging to datatype: wikibase:Math[41].

Some contributors have created property classes to help to describe Wikidata items belonging to different domains, especially for new contributors who often find it difficult to find the relevant properties to describe a given entity. Figure 14 shows a snapshot of the available property classes[42] and highlights lighthouse (Q39715).

Figure 15 shows the details of two such curated Wikidata items: lighthouse (Q39715)[43] and Wikidata property related to lighthouses (Q28739677)[44] and shows a screenshot of some of the properties proposed by them.

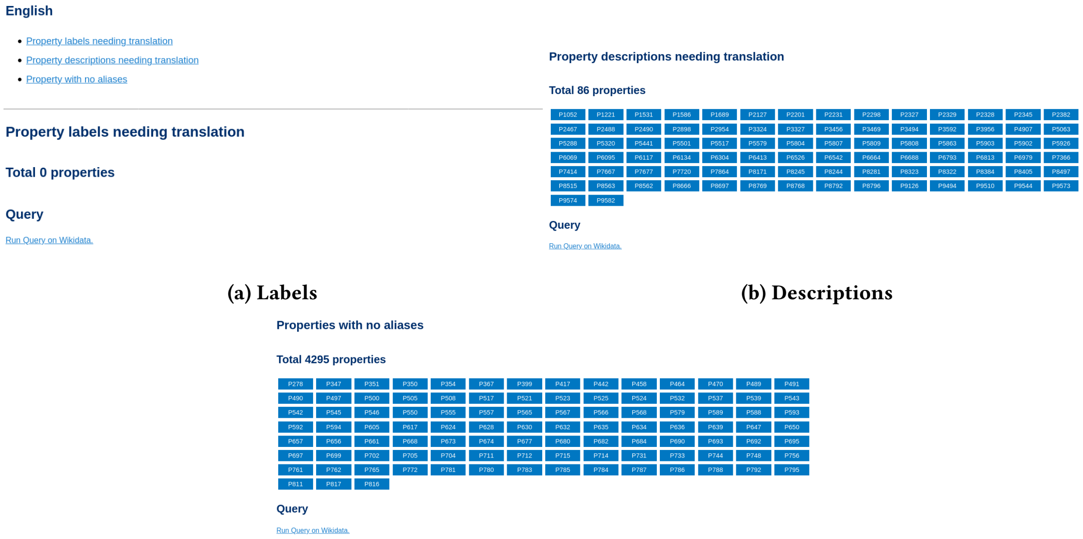
[40] https://wdprop.toolforge.org/datatypes.html

[41] https://wdprop.toolforge.org/datatype.html?datatype=wikibase:Math

[42] https://wdprop.toolforge.org/classes.html

[43] https://wdprop.toolforge.org/class.html?class=Q39715

[44] https://wdprop.toolforge.org/class.html?class=Q28739677



English

- Property labels needing translation
- Property descriptions needing translation
- Property with no aliases

Property labels needing translation

Total 0 properties

Query

Run Query on Wikidata.

(a) Labels

Property descriptions needing translation

Total 86 properties

P1052	P1221	P1531	P1586	P1689	P2127	P2201	P2231	P2298	P2327	P2329	P2328	P2345	P2382
P2467	P2488	P2490	P2898	P2954	P3324	P3327	P3456	P3469	P3494	P3592	P3956	P4907	P5063
P5288	P5320	P5441	P5501	P5517	P5579	P5804	P5807	P5809	P5808	P5863	P5903	P5902	P5926
P6069	P6095	P6117	P6134	P6304	P6413	P6526	P6542	P6664	P6688	P6793	P6813	P6979	P7366
P7414	P7667	P7677	P7720	P7864	P8171	P8245	P8244	P8281	P8323	P8322	P8384	P8405	P8497
P8515	P8563	P8562	P8666	P8697	P8769	P8768	P8792	P8796	P9126	P9494	P9510	P9544	P9573
P9574	P9582												

Query

Run Query on Wikidata.

(b) Descriptions

Properties with no aliases

Total 4295 properties

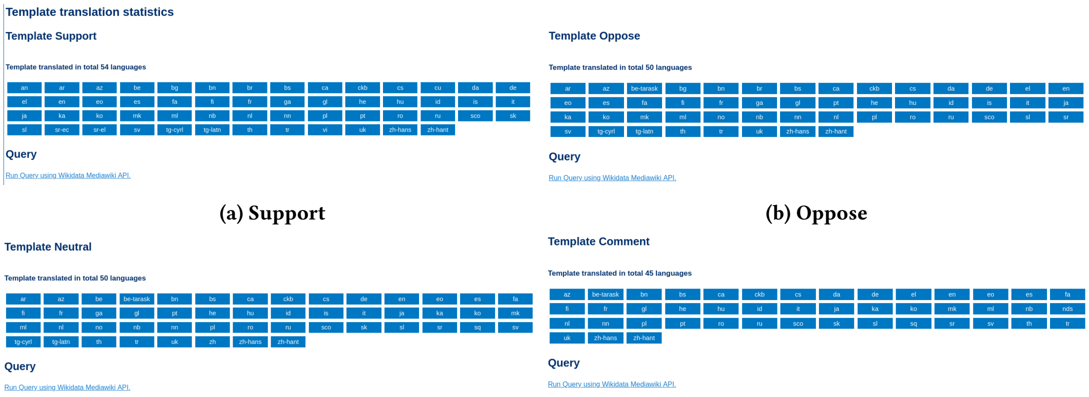
P278	P347	P351	P350	P354	P367	P399	P417	P442	P458	P464	P470	P489	P491
P490	P497	P500	P505	P508	P517	P521	P523	P525	P524	P532	P537	P539	P543
P542	P545	P546	P550	P555	P557	P565	P567	P566	P568	P579	P589	P588	P593
P592	P594	P605	P617	P624	P628	P630	P632	P635	P634	P636	P639	P647	P650
P657	P656	P661	P668	P673	P674	P677	P680	P682	P684	P690	P693	P692	P695
P697	P699	P702	P705	P704	P711	P712	P715	P714	P731	P733	P744	P748	P756
P761	P762	P765	P772	P781	P780	P783	P785	P784	P787	P786	P788	P792	P795
P811	P817	P816											

Query

Run Query on Wikidata.

(c) Aliases

Figure 11: Property information in the English (en) language



Template translation statistics

Template Support

Template translated in total 54 languages

an	ar	az	be	bg	bn	br	bs	ca	ckb	cs	cu	da	de
el	en	eo	es	fa	fi	fr	ga	gl	he	hu	id	is	it
ja	ka	ko	mk	ml	nb	nl	nn	pl	pt	ro	ru	sco	sk
sl	sr-ec	sr-el	sv	tg-cyrl	tg-latn	th	tr	vi	uk	zh-hans	zh-hant		

Query

Run Query using Wikidata Mediawiki API.

(a) Support

Template Oppose

Template translated in total 50 languages

ar	az	be-tarask	bg	bn	br	bs	ca	ckb	cs	da	de	el	en
eo	es	fa	fi	fr	ga	gl	pt	he	hu	id	is	it	ja
ka	ko	mk	ml	no	nb	nn	nl	pl	ro	ru	sco	sl	sr
sv	tg-cyrl	tg-latn	th	tr	uk	zh-hans	zh-hant						

Query

Run Query using Wikidata Mediawiki API.

(b) Oppose

Template Neutral

Template translated in total 50 languages

ar	az	be	be-tarask	bn	bs	ca	ckb	cs	de	en	eo	es	fa
fi	fr	ga	gl	pt	he	hu	id	is	it	ja	ka	ko	mk
ml	nl	no	nb	nn	pl	ro	ru	sco	sk	sl	sr	sq	sv
tg-cyrl	tg-latn	th	tr	uk	zh	zh-hans	zh-hant						

Query

Run Query using Wikidata Mediawiki API.

(c) Neutral

Template Comment

Template translated in total 45 languages

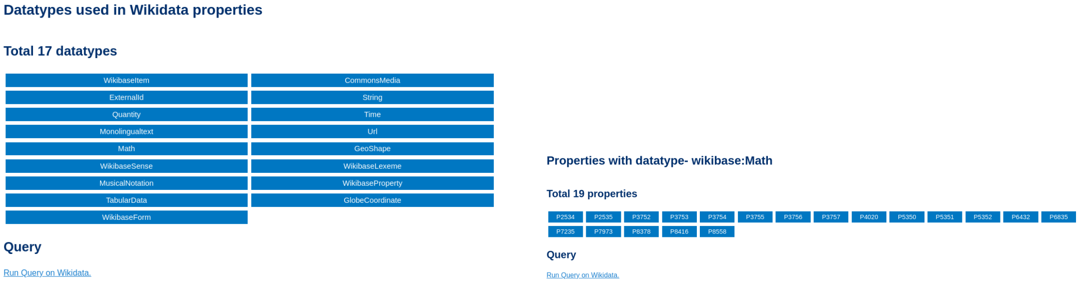
az	be-tarask	bn	bs	ca	ckb	cs	da	de	el	en	eo	es	fa
fi	fr	gl	he	hu	id	it	ja	ka	ko	mk	ml	nb	nds
nl	nn	pl	pt	ro	ru	sco	sk	sl	sq	sr	sv	th	tr
uk	zh-hans	zh-hant											

Query

Run Query using Wikidata Mediawiki API.

(d) Comment

Figure 12: Translation Statistics of templates used for property discussion: Support, Oppose, Neutral and Comment



Datatypes used in Wikidata properties

Total 17 datatypes

WikibaseItem	CommonsMedia
ExternalId	String
Quantity	Time
Monolingualtext	Url
Math	GeoShape
WikibaseSense	WikibaseLexeme
MusicalNotation	WikibaseProperty
TabularData	GlobeCoordinate
WikibaseForm	

Query

Run Query on Wikidata.

(a) Datatypes

Properties with datatype- wikibase:Math

Total 19 properties

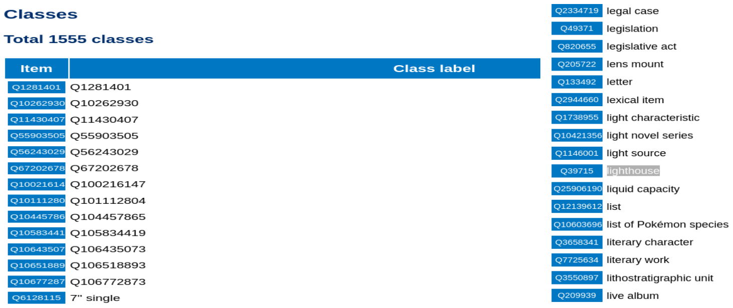
P2534	P2535	P3752	P3753	P3754	P3755	P3756	P3757	P4020	P5350	P5351	P5352	P6432	P6835
P7235	P7973	P8378	P8416	P8558									

Query

Run Query on Wikidata.

(b) Properties belonging to datatype: wikibase:Math

Figure 13: Datatypes



Classes

Total 1555 classes

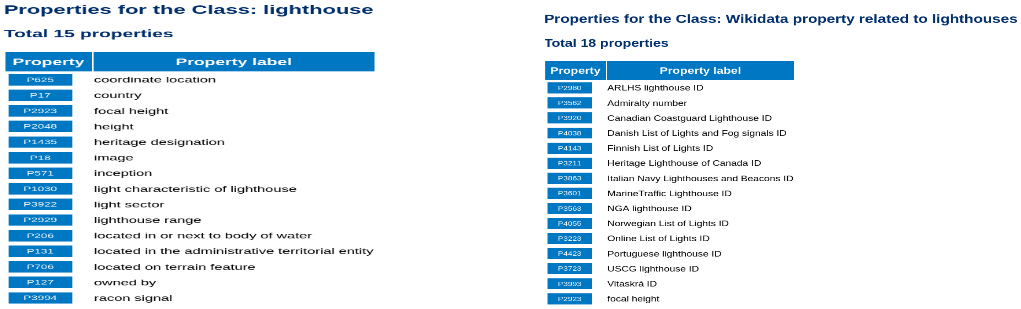
Item	Class label
Q1281401	Q1281401
Q10262930	Q10262930
Q11430407	Q11430407
Q55903505	Q55903505
Q56243029	Q56243029
Q67202678	Q67202678
Q10021614	Q100216147
Q10111280	Q101112804
Q10445786	Q104457865
Q10583441	Q105834419
Q10643507	Q106435073
Q10651889	Q106518893
Q10677287	Q106772873
Q6128115	7" single

(a) Property classes

Q2334719	legal case
Q49371	legislation
Q820655	legislative act
Q205722	lens mount
Q133492	letter
Q2944660	lexical item
Q1738955	light characteristic
Q10421356	light novel series
Q1146001	light source
Q39715	lighthouse
Q25906190	liquid capacity
Q12139612	list
Q10603696	list of Pokémon species
Q3658341	literary character
Q7725634	literary work
Q3550897	lithostratigraphic unit
Q209939	live album

(b) Highlighted property class: lighthouse (Q39715)

Figure 14: Property classes: curation of properties for a particular class



Properties for the Class: lighthouse

Total 15 properties

Property	Property label
P625	coordinate location
P17	country
P2923	focal height
P2048	height
P1435	heritage designation
P18	image
P571	inception
P1030	light characteristic of lighthouse
P3922	light sector
P2929	lighthouse range
P206	located in or next to body of water
P131	located in the administrative territorial entity
P706	located on terrain feature
P127	owned by
P3994	racon signal

(a) Properties curated on property class: lighthouse (Q39715)

Properties for the Class: Wikidata property related to lighthouses

Total 18 properties

Property	Property label
P2980	ARLHS lighthouse ID
P3562	Admiralty number
P3920	Canadian Coastguard Lighthouse ID
P4038	Danish List of Lights and Fog signals ID
P4143	Finnish List of Lights ID
P3211	Heritage Lighthouse of Canada ID
P3863	Italian Navy Lighthouses and Beacons ID
P3601	MarineTraffic Lighthouse ID
P3563	NGA lighthouse ID
P4055	Norwegian List of Lights ID
P3223	Online List of Lights ID
P4423	Portuguese lighthouse ID
P3723	USCG lighthouse ID
P3993	Vitaskrá ID
P2923	focal height

(b) Properties curated on property class: Wikidata property related to lighthouses (Q28739677)

Figure 15: Properties curated on two property classes

The curators may be interested to know the translation statistics of all the properties considered in a given class. Figure 16 shows the statistics of property translations of the two property classes.

The translation statistics of the properties curated and used on the WikiProjects can also be seen. Take, for example, Figure 17 shows the translation statistics of labels of the 37 properties used on WikiProject Museums[45].

In Figure 18, the translation path of a Wikidata property, P856 (official website)[46] is shown. This is an extension to the work previously done in [13], where a tabular column was used to highlight the translation of property labels, descriptions, and aliases. On clicking any property page (e.g., P17[47], it is possible to obtain the current translation statistics in different languages, links to the provenance information (number of references on statements describing the property, number of equivalent properties, usage of references in the statements using the property), translation path, and visualization of the translation path.

4.3 Discussion

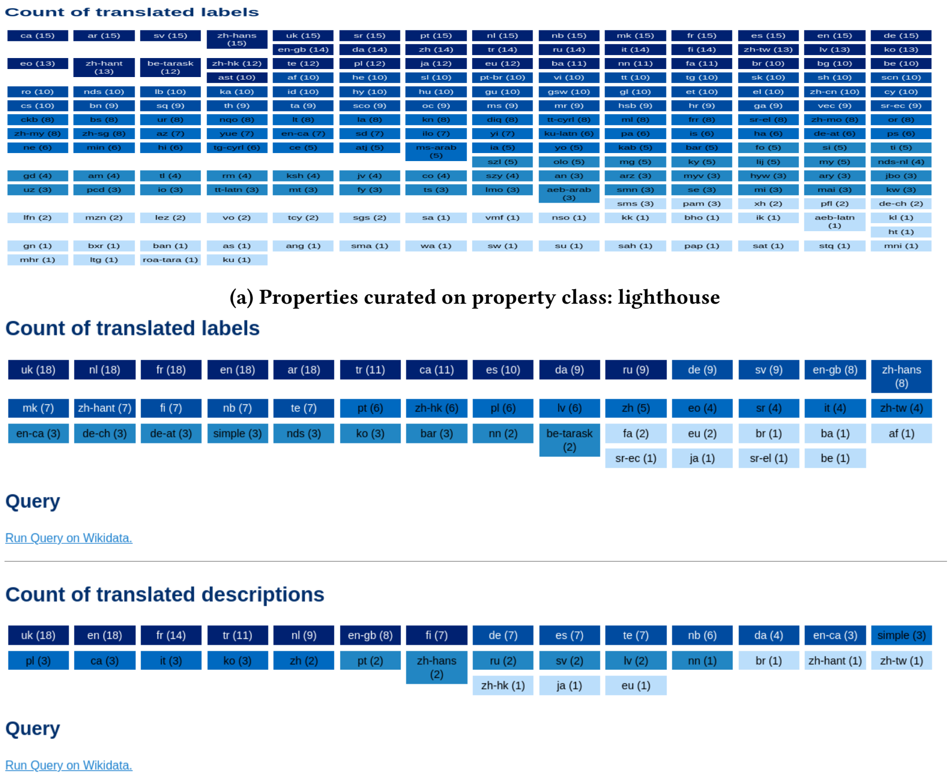
Wikidata is highly evolving, and new properties are being regularly proposed and translated. The WDProp interface has also undergone several changes during the last couple of years to incorporate these changes. Take, for example, WDProp initially displayed all the properties on the properties page [14] (less than 5000 properties at that time). But this number is currently around 9,000 properties. So now, the interface shows only the first 100 properties obtained from the SPARQL query. However, the user now has the SPARQL query link to obtain the complete list. The SPARQL query links and links to Wikidata Mediawiki API calls have now been integrated into all the pages on WDProp. This feature gives the researchers and practitioners the possibility to not only verify the visualization results but also modify these SPARQL/API queries based on their specific requirements.

Some commonly used properties also face vandalism. The commit history of some initial properties has therefore become long, which means obtaining the complete translation path as shown in Figure 18 takes a significant amount of time. Also, the number of property classes and WikiProjects is increasing, which means the current display of these lists may require some modifications in the future, especially the possibility to search and filter the desired information.

5 CONCLUSION AND FUTURE WORKS

With the growing usage of mobile devices and internet penetration across the world, web developers and content producers are looking for ways to ensure the availability of information in the local languages of the readers. Considering the number of languages, it is

[45] https://wdprop.toolforge.org/wikiproject.html?project=Wikidata:WikiProject%20Museums
[46] https://wdprop.toolforge.org/pathviz.html?property=P856
[47] https://wdprop.toolforge.org/property.html?property=P17



Count of translated labels

ca (15) ar (15) sv (15) zh-hans (15) uk (15) sr (15) pt (15) nl (15) nb (15) mk (15) fr (15) es (15) en (15) de (15)
en-gb (14) da (14) zh (14) tr (14) ru (14) it (14) fi (14) zh-tw (13) lv (13) ko (13)
eo (13) zh-hant (13) be-tarask (12) zh-hk (12) te (12) pl (12) ja (12) eu (12) ba (11) nn (11) fa (11) br (10) bg (10) be (10)
ast (10) af (10) he (10) sl (10) pt-br (10) vi (10) tt (10) tg (10) sk (10) sh (10) scn (10)
ro (10) nds (10) lb (10) ka (10) id (10) hy (10) hu (10) gu (10) gsw (10) gl (10) et (10) el (10) zh-cn (10) cy (10)
cs (10) bn (9) sq (9) th (9) ta (9) sco (9) oc (9) ms (9) mr (9) hsb (9) hr (9) ga (9) vec (9) sr-ec (9)
ckb (8) bs (8) ur (8) nqo (8) lt (8) la (8) kn (8) diq (8) tt-cyrl (8) ml (8) frr (8) sr-el (8) zh-mo (8) or (8)
zh-my (8) zh-sg (8) az (7) yue (7) en-ca (7) sd (7) ilo (7) yi (7) ku-latn (6) pa (6) is (6) ha (6) de-at (6) ps (6)
ne (6) min (6) hi (6) tg-cyrl (6) ce (5) atj (5) ms-arab (5) ia (5) yo (5) kab (5) bar (5) fo (5) si (5) ti (5)
szl (5) olo (5) mg (5) ky (5) lij (5) my (5) nds-nl (4)
gd (4) am (4) tl (4) rm (4) ksh (4) jv (4) co (4) szy (4) an (3) arz (3) myv (3) hyw (3) ary (3) jbo (3)
uz (3) pcd (3) io (3) tt-latn (3) mt (3) fy (3) ts (3) lmo (3) aeb-arab (3) smn (3) se (3) mi (3) mai (3) kw (3)
sms (3) pam (3) xh (2) pfl (2) de-ch (2)
lfn (2) mzn (2) lez (2) vo (2) tcy (2) sgs (2) sa (1) vmf (1) nso (1) kk (1) bho (1) ik (1) aeb-latn (1) kl (1)
ht (1)
gn (1) bxr (1) ban (1) as (1) ang (1) sma (1) wa (1) sw (1) su (1) sah (1) pap (1) sat (1) stq (1) mni (1)
mhr (1) ltg (1) roa-tara (1) ku (1)

(a) Properties curated on property class: lighthouse

Count of translated labels

uk (18) nl (18) fr (18) en (18) ar (18) tr (11) ca (11) es (10) da (9) ru (9) de (9) sv (9) en-gb (8) zh-hans (8)
mk (7) zh-hant (7) fi (7) nb (7) te (7) pt (6) zh-hk (6) pl (6) lv (6) zh (5) eo (4) sr (4) it (4) zh-tw (4)
en-ca (3) de-ch (3) de-at (3) simple (3) nds (3) ko (3) bar (3) nn (2) be-tarask (2) fa (2) eu (2) br (1) ba (1) af (1)
sr-ec (1) ja (1) sr-el (1) be (1)

Query

Run Query on Wikidata.

Count of translated descriptions

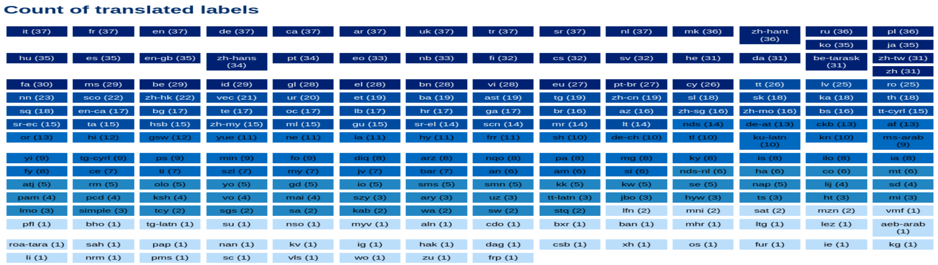
uk (18) en (18) fr (14) tr (11) nl (9) en-gb (8) fi (7) de (7) es (7) te (7) nb (6) da (4) en-ca (3) simple (3)
pl (3) ca (3) it (3) ko (3) zh (2) pt (2) zh-hans (2) ru (2) sv (2) lv (2) nn (1) br (1) zh-hant (1) zh-tw (1)
zh-hk (1) ja (1) eu (1)

Query

Run Query on Wikidata.

(b) Properties curated on property class: Wikidata property related to lighthouses

Figure 16: Statistics of property translations curated on two property classes



Count of translated labels

it (37) fr (37) en (37) de (37) ca (37) ar (37) uk (37) tr (37) sr (37) nl (37) mk (36) zh-hant (36) ru (36) pl (36)
ko (35) ja (35)
hu (35) es (35) en-gb (35) zh-hans (34) pt (34) eo (33) nb (33) fi (32) cs (32) sv (32) he (31) da (31) be-tarask (31) zh-tw (31)
zh (31)
fa (30) ms (29) be (29) id (29) gl (28) el (28) bn (28) vi (28) eu (27) pt-br (27) cy (26) tt (26) lv (25) ro (25)
nn (23) sco (22) zh-hk (22) vec (21) ur (20) et (19) ba (19) ast (19) tg (19) zh-cn (19) sl (18) sk (18) ka (18) th (18)
sq (18) en-ca (17) bg (17) te (17) oc (17) lb (17) hr (17) ga (17) br (16) az (16) zh-sg (16) zh-mo (16) bs (16) tt-cyrl (15)
sr-ec (15) ta (15) hsb (15) zh-my (15) ml (15) gu (15) sr-el (14) scn (14) mr (14) lt (14) nds (14) de-at (13) ckb (13) af (13)
or (13) hi (12) gsw (12) yue (11) ne (11) la (11) hy (11) frr (11) sh (10) de-ch (10) tl (10) ku-latn (10) kn (10) ms-arab (9)
yi (9) tg-cyrl (9) ps (9) min (9) fo (9) diq (8) arz (8) nqo (8) pa (8) mg (8) ky (8) is (8) ilo (8) ia (8)
fy (8) ce (7) ti (7) szl (7) my (7) jv (7) bar (7) an (6) am (6) si (6) nds-nl (6) ha (6) co (6) mt (6)
atj (5) rm (5) olo (5) yo (5) gd (5) io (5) sms (5) smn (5) kk (5) kw (5) se (5) nap (5) lij (4) sd (4)
pam (4) pcd (4) ksh (4) vo (4) mai (4) szy (3) ary (3) uz (3) tt-latn (3) jbo (3) hyw (3) ts (3) ht (3) mi (3)
lmo (3) simple (3) tcy (2) sgs (2) sa (2) kab (2) wa (2) sw (2) stq (2) lfn (2) mni (2) sat (2) mzn (2) vmf (1)
pfl (1) bho (1) tg-latn (1) su (1) nso (1) myv (1) aln (1) cdo (1) bxr (1) ban (1) mhr (1) ltg (1) lez (1) aeb-arab (1)
roa-tara (1) sah (1) pap (1) nan (1) kv (1) ig (1) hak (1) dag (1) csb (1) xh (1) os (1) fur (1) ie (1) kg (1)
li (1) nrm (1) pms (1) sc (1) vls (1) wo (1) zu (1) frp (1)

Figure 17: Statistics of property translations curated by Wikidata WikiProject Museums

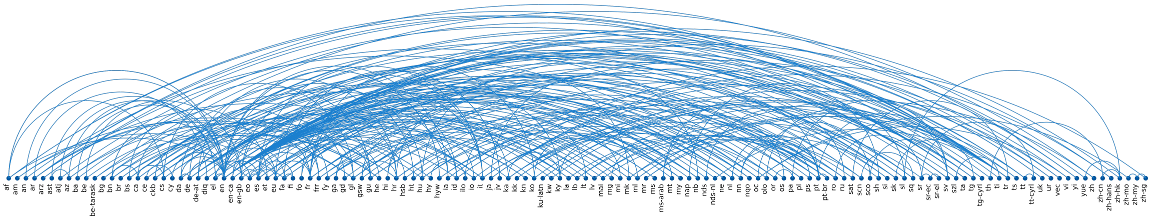


Figure 18: Translation of Wikidata property labels: official website (P856)

not easy for a few developers to produce multilingual information. Wikipedia and Wikidata have long shown how such multilingual information can be produced collaboratively and in an open manner. However, considering the disparity in the number of contributors in different languages, we find that not all languages are equally represented on these websites. This article focused on Wikidata properties and their translations and discussed the design and development of WDProp for this study. It also demonstrated how the web application can be used to perform granular analysis on Wikidata properties and their curation.

The future course of action includes options to download the data as CSV or JSON files and a multilingual user interface. Another possible direction is to explore how the property translation flow can be used to suggest the untranslated properties to the contributors of a given language. Some aspects of this work can also be integrated into Wikibase installations using SPARQL endpoints and Mediawiki API, but this requires additional experiments.

ACKNOWLEDGMENTS

The author would like to thank the Wikidata and Wikimedia community members for their feedback during Wikimedia Hackathon 2018, WikidataCon 2019, and ContribuLing 2021 for the feedback, thanks to which the web application has been improved and has helped to provide a written version of this article.

REFERENCES

[1] Erik Borra, David Laniado, Esther Weltevrede, Michele Mauri, Giovanni Magni, Tommaso Venturini, Paolo Ciuccarelli, Richard Rogers, and Andreas Kaltenbrunner. 2015. A Platform for Visually Exploring the Development of Wikipedia Articles. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*. AAAI Press, 711–712. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10565

[2] Thibaut Chamard and John Samuel. 2019. *Multilingual Wikidata Property Translation Flow Dataset.* https://doi.org/10.5281/zenodo.3271358

[3] Lars Christoph Gleim, Rafael Schimassek, Dominik Hüser, Maximilian Peters, Christoph Krämer, Michael Cochez, and Stefan Decker. 2020. SchemaTree: Maximum-Likelihood Property Recommendation for Wikidata. In *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings (Lecture Notes in Computer Science, Vol. 12123)*, Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez (Eds.). Springer, 179–195. https://doi.org/10.1007/978-3-030-49461-2_11

[4] Simon Gottschalk and Elena Demidova. 2017. MultiWiki: Interlingual Text Passage Alignment in Wikipedia. *TWEB* 11, 1 (2017), 6:1–6:30. http://doi.acm.org/10.1145/3004296

[5] Lucie-Aimée Kaffee, Kemele M. Endris, and Elena Simperl. 2019. When humans and machines collaborate: cross-lingual label editing in wikidata. In *Proceedings of the 15th International Symposium on Open Collaboration, OpenSym 2019, Skövde, Sweden, August 20-22, 2019*, Björn Lundell, Jonas Gamalielsson, Lorraine Morgan, and Gregorio Robles (Eds.). ACM, 16:1–16:9. https://doi.org/10.1145/3306446.3340826

[6] Lucie-Aimée Kaffee, Alessandro Piscopo, Pavlos Vougiouklis, Elena Simperl, Leslie Carr, and Lydia Pintscher. 2017. A Glimpse into Babel: An Analysis of Multilinguality in Wikidata. In *Proceedings of the 13th International Symposium on Open Collaboration (OpenSym '17)*. Association for Computing Machinery, Galway, Ireland, 1–5. https://doi.org/10.1145/3125433.3125465

[7] Timothy Kanke. 2018. Preliminary Exploration of Knowledge Curation Activities in Wikidata WikiProjects. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. ACM, Fort Worth Texas USA, 349–350. https://doi.org/10.1145/3197026.3203878

[8] Timothy Kanke. 2020. Knowledge curation work in Wikidata WikiProject discussions. *Library Hi Tech* 39, 1 (Jan. 2020), 64–79. https://doi.org/10.1108/LHT-04-2019-0087

[9] Dustin Lange, Christoph Böhm, and Felix Naumann. 2010. Extracting structured information from Wikipedia articles to populate infoboxes. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. ACM Press, Toronto, ON, Canada, 1661. https://doi.org/10.1145/1871437.1871698

[10] Michael Luggen, Julien Audiffren, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2021. Wiki2Prop: A Multimodal Approach for Predicting Wikidata Properties from Wikipedia. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 2357–2366. https://doi.org/10.1145/3442381.3450082

[11] Claudia Müller-Birn, Benjamin Karran, Janette Lehmann, and Markus Luczak-Rösch. 2015. Peer-production system or collaborative ontology engineering effort: what is Wikidata?. In *Proceedings of the 11th International Symposium on Open Collaboration (OpenSym '15)*. Association for Computing Machinery, San Francisco, California, 1–10. https://doi.org/10.1145/2788993.2789836

[12] John Samuel. 2017. Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata. In *Metadata and Semantic Research (Communications in Computer and Information Science)*, Emmanouel Garoufallou, Sirje Virkus, Rania Siatri, and Damiana Koutsomiha (Eds.). Springer International Publishing, Cham, 167–172. https://doi.org/10.1007/978-3-319-70863-8_16

[13] John Samuel. 2018. Analyzing and Visualizing Translation Patterns of Wikidata Properties. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (Lecture Notes in Computer Science)*, Patrice Bellot, Chiraz Trabelsi, Josiane Mothe, Fionn Murtagh, Jian Yun Nie, Laure Soulier, Eric SanJuan, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 128–134. https://doi.org/10.1007/978-3-319-98932-7_12

[14] John Samuel. 2018. Towards Understanding and Improving Multilingual Collaborative Ontology Development in Wikidata. In *Wikidata Workshop 2018*. http://wikiworkshop.org/2018/papers/wikiworkshop2018_paper_12.pdf

[15] Moritz Stefaner, Dario Taraborelli, and Giovani L Ciampaglia. 2011. Notabilia–Visualizing Deletion Discussions on Wikipedia. http://well-formed-data.net/archives/570/notabilia-visualizing-deletion-discussions-on-wikipedia

[16] Thomas Pellissier Tanon and Lucie-Aimée Kaffee. 2018. Property Label Stability in Wikidata: Evolution and Convergence of Schemas in Collaborative Knowledge Bases. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*. ACM, 1801–1803. http://doi.acm.org/10.1145/3184558.3191643

[17] Denny Vrandečić. 2012. Wikidata: a new platform for collaborative data collection. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*. ACM Press, Lyon, France, 1063. https://doi.org/10.1145/2187980.2188242

[18] Eva Zangerle, Wolfgang Gassler, Martin Pichl, Stefan Steinhauser, and Günther Specht. 2016. An Empirical Evaluation of Property Recommender Systems for Wikidata and Collaborative Knowledge Bases. In *Proceedings of the 12th International Symposium on Open Collaboration (OpenSym '16)*. Association for Computing Machinery, Berlin, Germany, 1–8. https://doi.org/10.1145/2957792.2957804

Group Formation in a Cross-Classroom Collaborative Project-Based Learning Environment

Gail L. Rolle-Greenidge
The University of the West Indies, Cave Hill Campus
gail.rolle@mycavehill.uwi.edu

Paul A. Walcott*
The University of the West Indies, Cave Hill Campus
paul.walcott@cavehill.uwi.edu

ABSTRACT

Cross-Classroom Collaborative Project-Based Learning (C3PjBL) requires the formation of project-groups by pairing student-groups across classrooms. Unfortunately, due to the configuration of these groups, the group formation techniques found in the literature are unable to automatically create project-groups for C3PjBL. This paper describes an automatic project-group formation technique for C3PjBL which utilizes clustering to create homogeneous student-groups, based on the students' perceived technological and higher-order thinking skills (student characteristics). Student-groups, from different classrooms, are then paired using an optimization technique to form project-groups. In our results, we present a comparison of the performance of a random group formation technique and our technique. We observed that automatic group formation using an n-dimensional space of student characteristics and k-means clustering is more effective than random group formation and, the strategy of forming homogeneous student-groups and heterogeneous project-group for C3PjBL creates more compatible group compositions than random grouping.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing**; • **Collaborative and social computing systems and tools**;

KEYWORDS

Collaborative and social computing, Group formation, Clustering, Cross-classroom collaborative project-based learning, C3PjBL, G2Group

ACM Reference Format:
Gail L. Rolle-Greenidge and Paul A. Walcott. 2021. Group Formation in a Cross-Classroom Collaborative Project-Based Learning Environment. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3479986.3479997

1 INTRODUCTION

Cross-classroom collaborative project-based learning (C3PjBL) is a teaching and learning strategy that requires student-groups from one classroom to work collaboratively with student-groups from another to complete curriculum-based activities using wikis [1, 20]. Before collaboration can occur however, a project-group must be formed by pairing a student-group from one classroom with a student-group from another. Although manual group formation is possible in this environment, it is time-consuming and error prone. Due to this complexity, group formation in C3PjBL needs to be automated if teachers are expected to effectively work in this environment [1].

Recently, factors that impact automatic group formation in online collaborative environments have been the focus of several researchers. These factors include: students' characteristics and learning attributes [2]; optimization for group formation [3]; and, attributes of group formation and grouping techniques [4]. Through the process of selecting students to participate in a group, teachers can combine several factors to form collaborative groups that will foster meaningful interactions and desired learning outcomes [4]. However, creating these complex groups often necessitates computational backing to be successful [5]. Two automatic group creation approaches are presented in the literature: the random selection method and computational techniques (namely algorithms). Although the random selection method has been preferred for online group formation due to its simplicity, Cruz and Isotani [6], indicated that randomly created groups often pose challenges such as "disproportional participation of individuals, demotivation, and resistance to group work in futures [sic] activities". [2], also added that groups that are randomly created could result in groups being homogenous instead of heterogeneous.

To the best of our knowledge, no existing open project-based learning systems support cross-classroom collaboration and project-group creation for C3PjBL. In this paper, we describe an automatic project-group formation technique for C3PjBL which utilizes clustering to create homogeneous groups, based on students' perceived technological and higher-order thinking skills [7]. Homogeneous student-groups, from different classrooms are then paired using an optimization technique to form project-groups. We present a comparison of the performance of a random group formation technique and our computational technique.

Our research expands the current research on automatic group formation and brings to the fore effective heterogeneous and homogenous grouping for C3PjBL based on students' technological and higher-order thinking skills. Combining these skills to group students creates a balance in the groups which can improve group performance and create a positive synergy among group members, thus improving their higher-order thinking skills.

*Corresponding author

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479997

Without automatic group formation in C3PjBL, teachers may be unwilling to engage in C3PjBL [1] thus not benefiting from Collaborative PjBL. Also, manual group formation is tedious and time-consuming and may lead to incompatible groups.

In the next section of the paper, the research literature is reviewed, followed by a discussion of the novelty of the proposed approach, a description of the present study, the results, discussion and conclusion.

2 BACKGROUND

2.1 Group formation

Group formation in online collaboration can be either manual or automatic [9]. Manual group formation can be achieved either by self-selection or instructor assigned. The self-selection approach does not guarantee a balanced group, since learners choose to join a group that is most suitable for them. This may lead to a less than ideal group [10]. Although the instructor assigned grouping technique guarantees a more balanced grouping, it becomes complex when large number of students must be grouped manually [11].

Automatic group formation provides the option of group creation with or without instructor input [12]. Random selection and computational techniques (for example, algorithms) are two techniques used to achieve automatic group formation. Random group formation is the approach most frequently used by instructors because of its simplified implementation, whereby social and academic heterogeneity can possibly be achieved given that students have an equal opportunity of being a member of any group [2]. The authors further point out that although this grouping method is popular in learning management systems, such as Moodle, the level of heterogeneity may not match the diversity in learning capabilities which is required for effective grouping. Authors like Maqtary and his colleagues [4] caution that randomly forming groups for collaboration often results in a mismatch of students' skills and characteristics, and a group composition that poorly represents the structure of successful groups [9].

Group formation in the form of heterogenous and homogeneous groups is a topic of interest in recent research. A study was conducted by Wichmann et al. [8] to determine whether groups formed based on learner behavior impacted productivity when students who were classified as either high, average, or low-level were randomly assigned to heterogeneous or homogenous groups.

2.2 Clustering

Clustering is used to find groups of objects with related characteristics [2]. Romero et al [13] defined clustering based on the premise of maximizing the similarity among the object groups in a cluster while minimizing the similarity between the object groups in different clusters. In online collaborative learning, clustering has been used to group students according to their collaboration competence level [2], predict students' academic performance [14] and group students to give them differentiated guidance according to their learning skills and other characteristics [15].

Maina et al. [2], in their work, applied an intelligent grouping clustering algorithm to automatically form heterogeneous groups using students' collaboration competence levels. Similar work has also been done by Valetts and Gesa [14], however, they proposed a different clustering method to group students using their collaboration competences. In another work, Tang and McCalla [16], employed a clustering technique to group students with similar learning characteristics to promote group-based collaborative learning. Anaya and Boticario [15] also applied a clustering algorithm to group students according to their collaboration level (high, low, or medium) to evaluate student interactions.

2.2.1 *K-means clustering.*

K-means is an unsupervised learning algorithm used for clustering. K-means clustering works by partitioning "n" objects into k clusters in which each object belongs to the cluster with the closest mean [17], thus, each cluster formed is associated with a centroid. The K-means algorithm minimizes the sum of distances between the points and their respective cluster centroid [18]. Drake and Gyimah [19] developed a simple algorithm (Algorithm 1) to perform K-means clustering which can be extended or modified to implement user characteristics and attributes for group formation [19].

Algorithm 1 K-means Clustering Algorithm

```
a. Clusters the data into $k$ groups where $k$ is predefined.
b. Select $k$ points at random as cluster centers.
c. Assign objects to their closest cluster center according to the
   *Euclidean distance* function.
d. Calculate the centroid or mean of all objects in each cluster.
e. These centroids form the new cluster centers.
f. Repeat steps c, d, and e until the same points are assigned to
   each cluster in consecutive rounds.
```

3 NOVELTY OF GROUP FORMATION FOR C3PJBL IN G2GCOLLABORATE

G2GCollaborate is a web-based platform that implements the C3PjBL environment [1]. It provides a robust method for the creation of institutions which are managed by institutional administrators (IAs). These administrators are responsible for the creation of project originators (POs) and project coordinators (PCs). POs create projects and are responsible for project-group formation, while the PCs create student-groups and guide them through the project. To encourage collaboration and the growth of a learning community, G2GCollaborate features: project creation; automatic student-group, and project-group formation; wikis (G2GWiki); user profiles; messaging; notifications; scaffolding (through project-roles and wiki templates); and, wiki publishing and search through a RESTful API [1, 20].

In G2GCollaborate, student-groups are formed in the classrooms participating in the C3PjBL project. Students are grouped based on their perceived technological skills and their higher-order thinking (HOT) skills. Once the student-groups are created, an optimization technique is employed to create project-groups by pairing student-groups from the participating classrooms. There are two main approaches to group formation in G2GCollaborate: instructor group selection (that is, groups created manually by the project coordinators (PCs) and project originators (POs)) and automatic group creation using an extended K-means clustering algorithm.

4 PRESENT STUDY

In C3PjBL, an environment is created that promotes collaboration within student- and project-groups. Given that the group formation approach can impact the effectiveness of the group, C3PjBL has drawn on the research literature to inform the approach used. Wichmann et al. [8], noting that heterogeneous group composition is beneficial for learning in small-group tasks, studied 120 students placed in 29 small groups. These students were classified as low-, average- or high-level based on the number of characters they contributed to an essay assignment. They concluded that high-level students were more productive in heterogeneous groups, low-level students were more productive in homogeneous groups (since social loafing was reduced) and, overall, heterogeneous groups were better for learning communities. As such, C3PjBL forms small, homogeneous student-groups in each classroom to ensure that low-level students are grouped together; and, heterogeneous project-groups to ensure that high-level students are productive.

In the present study, we explore the efficiency of G2GCollaborate's group formation approach using a Social Studies research project designed for Caribbean Class 3 primary school students, based on their technological skills and Social Studies knowledge. The objective of the project was for students to create a wiki that discussed the construction of canoes by indigenous Caribbean people (of note however, due to the devastation left by a Category 5 hurricane, the project was never completed). The experiment presented here compares whether randomly formed groups or groups formed using G2GCollaborate automatic group formation approach, produced the more desirable group compositions [8].

4.1 Technology and knowledge skills

The technology and knowledge skills of 330 Class 3 students were determined during a survey conducted at nine Caribbean primary schools. The Instruments used were a Social Studies test (a national assessment) and a technology skills survey which queried students' perceptions of their ability to complete 12 technology skills (students indicated that they could either complete the skill or not). These technical skills included: searching for files on a computer, searching the Internet for information, creating web pages, downloading files, and uploading files. These instruments were administered by the class teachers and one of the researchers during school hours.

Of these 330 students, data from N=60 were selected for use in this experiment. These data represented N=26 students from a rural primary school and N=34 from an urban primary school. Parental consent was received from all participants.

4.2 Group formation

In C3PjBL, student characteristics are captured using an n-dimensional space. In the present experiment, a two-dimensional space (Search Skills, Social Studies Knowledge) was employed based on Social Studies Knowledge and Search Skills. The Search Skill dimension was created using a combination of the "searching for files on a computer" and "searching the Internet for information" survey items, converted to a percentage; and, the score on the Social Studies assessment, which was also converted to a percentage, was used as the Social Studies Knowledge dimension. These values were scaled to create a (1000,1000) two-dimensional space.

Since C3PjBL utilizes an n-dimensional space for student characteristics, if the behavior of learners was also recorded using learning analytics (for example), then these behaviors would simply become dimensions in the n-dimensional space.

K-means clustering [19] was used to cluster students into student-groups. For C3PjBL, two such student-groups needed to be created from collaborating classes and joined to create a project-group. An optimization technique was used to join student-groups from Class A (for example) to student-groups from Class B so that the most dissimilar student-group pairs were joined together. Essentially, this technique created heterogeneous project-groups as advocated by [13].

4.3 Comparison of group formation approaches

To determine the efficacy of the automatic group formation technique, we created N=6 student-groups each for Class A and Class B by randomly assigning students to each of the groups. Class A comprised N=34 students, while Class B comprised N=26 students. Next, for each student-group in Class A, we randomly selected a student-group in Class B to join it to, thus creating six project-groups. We then used the automatic group formation feature (called G2Group) in G2GCollaborate to create six student-groups (each) in Class A and Class B and the six project-groups and compared the results of the two approaches.

To determine which approach produced the better groupings, the Euclidean distance between the centroids of the two student-groups comprising the project-group were calculated for each project-group. The approach that maximized the sum of these distances was deemed better given that this ensured group heterogeneity.

5 RESULTS

The sizes of the student-groups generated by the Random and G2Group group formation approaches are presented in Table 1. Of note, the sizes of the student-groups created by G2Group vary from, as low as, one student (Group 1 in Class B; and, Group 6 in Class A) to, as high as, 18 students (Group 3 in Class A).

Table 2 shows the centroid (and Standard Deviation (SD)) in the two-dimensional space (Search Skills, Social Studies Knowledge) of each student-group in Class A and Class B, using the Random and G2Group group formation approaches. Student-group centroids with small SDs indicate closely packed clusters (for example, Group 5, Class B for the G2Group approach). The size of this two-dimensional space is (1000,1000) as described in Section 4.2.

Table 3 illustrates the project-group pairings from Class A and Class B with the associated Euclidean distances between student-group centroids, using both approaches.

6 DISCUSSION

The comparatively smaller standard deviations for the student-groups created using the automatic grouping technique of G2Group indicate that these groups are homogeneous (Table 2). Conversely, the relatively large standard deviations for the randomly created student-groups suggest non-homogeneous groupings. Further, the

Table 1: Group Sizes for Student-Groups in Class A and Class B for both Group Formation Approaches

Group	Random Group Formation Approach		G2Group Group Formation Approach	
	Class A	Class B	Class A	Class B
1	6	5	3	1
2	6	5	5	5
3	6	4	18	9
4	6	4	2	6
5	5	4	5	3
6	5	4	1	2
Total	**34**	**26**	**34**	**26**

Table 2: The Characteristics of the Student-Groups Created for Class A and Class B for both Group Formation Approaches

	Random Group Formation Approach(Search Skills, Social Studies Knowledge)				G2Group Group Formation Approach(Search Skills, Social Studies Knowledge)			
	Class A		Class B		Class A		Class B	
Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	(733.33, 583.33)	(413.12, 220.61)	(456, 610)	(448.00, 149.67)	(40.00, 750.00)	(0.00, 50.00)	(1000.00, 841.67)	(0.00, 78.62)
2	(680.00, 658.33)	(495.74, 139.34)	(520, 630)	(391.92, 169.12)	(1000.00, 744.44)	(0.00, 81.46)	(60.00, 462.50)	(52.92, 138.63)
3	(840.00, 675.00)	(391.92, 147.48)	(320, 613)	(397.99, 96.01)	(40.00, 362.50)	(0.00, 96.01)	(1000.00, 483.33)	(0.00, 94.28)
4	(840.00, 683.33)	(391.92, 116.90)	(600, 625)	(400.00, 225.00)	(200.00, 562.50)	(0.00, 73.95)	(200.00, 533.33)	(0.00, 62.36)
5	(296.00, 510.00)	(401.60, 267.86)	(520, 488)	(480.00, 170.93)	(1000.00, 460.00)	(0.00, 96.95)	(1000.00, 633.33)	(0.00, 23.57)
6	(840.00, 710.00)	(357.77, 65.19)	(760, 675)	(415.69, 182.00)	(200.00, 800.00)	(0.00, 0.00)	(466.67, 700.00)	(377.12, 40.82)

Table 3: Project-Group Pairings with Associated Euclidean Distances between Student-Group Centroids

Random Group Formation Approach			G2Group Group Formation Approach		
Group in Class A	Group in Class B	Euclidean Distance between Student-Group Centroids	Group in Class A	Group in Class B	Euclidean Distance between Student-Group Centroids
Group 6	Group 5	389.75	Group 1	Group 3	996.35
Group 1	Group 6	95.47	Group 2	Group 2	981.37
Group 2	Group 3	362.91	Group 3	Group 1	1072.94
Group 4	Group 1	390.94	Group 4	Group 5	803.13
Group 3	Group 4	245.15	Group 5	Group 6	584.85
Group 5	Group 2	254.12	Group 6	Group 4	700.00
Mean		**289.72**			**856.44**

position of each cluster in the (Search Skills, Social Studies Knowledge) space determines if the students would be considered high-level or low-level according to [8]. Student-groups occupying the top right portion of the (Search Skills, Social Studies Knowledge) space, comprise of students who are both skillful in the search process and have tested well on the Social Studies assessment; for example, Group 1 in Class B (with a centroid of (1000.00, 841.67))(Table 2) which was generated by G2Group.

Since high-level students are best placed in heterogeneous groups [8], G2Group pairs heterogeneous student-groups, from Class A and B, into project-groups. An example of this pairing occurred between Group 3 in Class A (40.00, 362.50) and Group 1 in Class B (1000.00, 841.67) (Table 3). The Euclidean distance between their centroids in the (Search Skills, Social Studies Knowledge) space of 1072.94, demonstrates their heterogeneity.

Overall, G2Group was much better at creating heterogeneous project-groups than the Random Grouping technique, based on the data presented in Table 3. The average Euclidean distance between the students-groups that were paired to form project-groups was only 289.72 for the Random Grouping technique, whereas G2Group's average Euclidean distance was 856.44. As argued in [2], the problem with the random-based approach is that homogenous groups might be created instead of heterogeneous groups (or vice-a-versa). Clearly, this was the case in our experiment.

Clustering using the standard k-means approach [19] can lead to peculiar group formations, such as groups containing one student (for example, Group 1 in Class B for G2Group (Table 1)) and very large groups (for example, Group 3 in Class A for G2Group (Table 1)) due to large numbers of students having similar technology and knowledge skills. PCs need to determine the desirability of these variations in group sizes before performing automatic group formation. For example, a PC might decide that a one-member student-group is acceptable since this group will always be joined with another student-group to form a project-group. Alternatively, PCs may choose to manually sub-divide larger groups. Another possibility is to modify the k-means approach so that it generates equal-sized groups.

The survey-based approach of acquiring student characteristics for group formation is an appropriate method for C3PjBL given that groups must be formed prior to the start of a C3PjBL activity. The presented technique of clustering in a (Search Skills, Social Studies Knowledge) space also provides great flexibility in group formation. For example, we chose the "search" technology skills from our survey data to perform group formation given that the C3PjBL activity was based on a Social Studies research project. Just as easily we could have selected word processing or communication skills as the basis for group formation. This feature makes G2Group a powerful automatic group formation tool.

The limitation of the study is that the approach used was not evaluated using control and treatment groups, however it was validated against state-of-the-art research.

7 CONCLUSION

This paper describes a novel group formation approach used in C3PjBL environments. Unlike classical group formation, C3PjBL requires both student-groups within classrooms to be created along with a project-group which is a pairing of student-groups across dispersed classrooms.

This group formation approach uses a k-means algorithm to create homogeneous student-groups based on clustering in an n-dimensional space (a two-dimensional (Search Skills, Social Studies Knowledge) space was used in the experiment presented). Subsequently, heterogeneous project-groups are created by pairing student-groups from two classrooms so that most dissimilar groups will be paired together.

The results of an experiment that compared the efficacy of a random group formation approach and the novel approach described in the present study revealed that: (1) automatic group formation using an n-dimensional space of student characteristics and k-means clustering is more effective than random group formation; (2) The strategy of forming homogeneous student-groups and heterogeneous project-group creates more compatible group compositions than random grouping; and, (3) A modified k-means clustering approach, which ensures equal group sizes, is desirable since it will ensure that student-group sizes are kept small.

Further work will explore the efficacy of creating same-sized groups versus variable-sized groups and their applicability to C3PjBL.

REFERENCES

[1] Paul. A. Walcott and Gail Rolle-Greenidge. 2021. A Cross-Classroom Collaborative Project-Based Learning Management System. Proceedings of EdMedia and Innovate Learning 2021—World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE), Online, 33-38.

[2] Elizaphan Maina, Robert Oboko, and Peter Waigo. 2017.Using machine learning techniques to support group formation in an online collaborative learning environment. International Journal of Intelligent Systems and Applications 9, 3 (March 2017), 26-33.https://doi.org/10.5815/ijisa.2017.03.04

[3] Hamid Sadeghi and Ahmad. A. Kardan. 2014. Toward effective group formation in computer-supported collaborative learning. Interactive Learning Environments24,3 (June 2014), 382–395. https://doi.org/10.1080/10494820.2013.851090.

[4] Naseebah Maqtary, Abdulqader Mohsen, and Kamal Bechkoum. 2017. Group formation techniques in computer-supported collaborative learning: A systematic literature review. Technology, Knowledge and Learning 24, 169-190.https://doi.org/10.1007/s10758-017-9332-1.

[5] Chinasa Odo, Judith Masthoff, and Nigel Beacham.2019. Group formation for collaborative learning. In: Isotani S., Millán E., Ogan A., Hastings P., McLaren B., Luckin R. (eds) Artificial Intelligence in Education. AIED 2019. Lecture Notes in Computer Science, vol 11626, 206-212. Springer, Cham. https://doi.org/10.1007/978-3-030-23207-8_39

[6] Wilmax. M. Cruz and SeijiIsotani. 2014. Group formation algorithms in collaborative learning contexts: A systematic mapping of the literature. Collaboration and Technology.Lecture Notes in Computer Science, vol 8658,199–214. https://doi.org/10.1007/978-3-319-10166-8_18.

[7] Paul. A. Walcott and Gail Rolle, 2014. Investigating the ICT competencies and technology access of Indigenous Dominican primary school students. In Proceedings of EdMedia 2014—World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE), Finland, 948-953.

[8] Astrid Wichmann, Tobias Hecking, Malte Elson, Nina Christmann, Thomas Herrmann, and H. Ulrich Hoppe. 2016. Group formation for small-group learning: Are heterogeneous groups more productive? In Proceedings of the 12th International Symposium on Open Collaboration (OpenSym '16). Association for Computing Machinery, New York, NY, USA, Article 14, 1–4. https://doi.org/10.1145/2957792.2965662

[9] Ivan Syrba and Maria Bielikova. 2015. Dynamic group formation as an approach to collaborative learning support. IEEE Transactions on Learning Technologies8, 2 (April-June 2015), 173–186. https://doi.org/10.1109/TLT.2014.2373374

[10] Zhillin Zheng and Niels Pinkwart. 2014. A discrete particle swarm optimization approach to compose heterogeneous learning groups. In proceedings of IEEE 14th international conference on advanced learning technologies, IEEE, Athens, Greece, 49–51. https://doi.org//10.1109/ICALT.2014.24

[11] Amir Mujkanovic, David Lowe, Kieth Willey, and Christian Guetl. 2012. Unsupervised learning algorithm for adaptive group formation: Collaborative learning support in remotely accessible laboratories. In Proceedings of International conference on information society (i-Society), IEEE, London, UK, 50–57.

[12] Sofiane Amara, Joaquin Macedo, Fatima Bendella, and Alexandre Santos. 2016. Group formation in mobile computer supported collaborative learning contexts: A systematic literature review. Journal of Educational Technology & Society 19, 2 (July 2015), 258–273.

[13] Cristóbal Romero, Sebastián Ventura, and Enrique García. Data mining in course management systems: Moodle case study and tutorial. Computers & Education 51, 1 (August 2008), 368-384. https://doi.org/10.1016/j.compedu.2007.05.016

[14] Laura. M. Valetts, Silvia. B. Navarro, and Ramón. F. Gesa. Modelling collaborative competence level using machine learning techniques. E-Learning. 56-60.

[15] Anaya. R. Antonio and Jesús. G. Boticario. 2011. Application of machine learning techniques to analyse student interactions and improve the collaboration process. Expert Systems with Applications 38,2, 1171-1181. https://doi.org/10.1016/j.eswa.2010.05.010.

[16] Tiffany. Y. Tangand Gordon McCalla. 2003. Smart recommendation for an evolving e-learning system. In Proceedings of Workshop on Technologies for Electronic

Documents for Supporting Learning, International Conference on Artificial Intelligence in Education, 699-710.

[17] Hedayetul. I. Shovonand Haque Mahfuza. 2012. An approach of improving students academic performance by using k means clustering algorithm and decision tree. International Journal of Advanced Computer Science and Applications 3, 8, 46-149.

[18] Qi Jianpeng, Yanwei Yu, Lihong Wang, Jinglei Liu, and Yingjie Wang. 2017. An effective and efficient hierarchical k-means clustering algorithm. International Journal of Distributed Sensor Networks 13, 8, 1-17. https://doi.org/10.1177/1550147717728627

[19] Delali. K. Dake and Ester Gyimah. 2019. Using k-means to determine learner typologies for project-based learning: a case study of the University of Education, Winneba. International Journal of Computer Application 178, 43, 29-34. https://doi.org/105120/ijca2019919320.

[20] Gail Rolle-Greenidge and Paul. A. Walcott. 2021. Design and Evaluation of a Wiki for Cross-Classroom Collaborative Project-Based Learning. Proceedings of EdMedia and Innovate Learning 2021—World Conference on Educational Media and Technology. Association for the Advancement of Computing in Education (AACE), Online, 27-32.

A Reference Model for Outside-in Open Innovation Platforms

Pablo Cruz
pcruz@inf.utfsm.cl
Universidad Técnica Federico Santa María
Valparaíso, Chile

Felipe Beroíza
felipe.beroiza@sansano.usm.cl
Universidad Técnica Federico Santa María
Valparaíso, Chile

Francisco Ponce
francisco.ponceme@sansano.usm.cl
Universidad Técnica Federico Santa María
Valparaíso, Chile

Hernán Astudillo
herman@inf.utfsm.cl
Universidad Técnica Federico Santa María
Valparaíso, Chile

ABSTRACT

The Open Innovation paradigm has spread widely since 2003, and led to the emergence of Open Innovation Platforms as software systems aiming at supporting and facilitating open innovation initiatives and projects. This software domain has matured up to a point where many functional concepts became notably common and used in these platforms. When implementing open innovation platforms, related people often struggle when defining expected functional characteristics due to the general application of the paradigm, making necessary the existence of a model that provide a set of potential functional features expected in the creation and development of this type of platform. Reference models provides a domain-specific set of clearly defined entities aiming at encouraging better communication in the domain. We propose in this paper a reference model for capturing and defining the functional features that could be implemented in outside-in oriented open innovation platforms. For building this reference model, we reviewed some of the already published reports of open innovation platforms implementations in order determine and define the potential functional features expected in this kind of platforms. We believe this knowledge base could ease software development and deployment decisions, especially at early stages where open innovation platforms adopters face development in a domain that as of this writing is still new to many people.

CCS CONCEPTS

• **Applied computing → Enterprise computing**.

KEYWORDS

open innovation, open innovation platforms, reference model

ACM Reference Format:
Pablo Cruz, Felipe Beroíza, Francisco Ponce, and Hernán Astudillo. 2021. A Reference Model for Outside-in Open Innovation Platforms. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3479986.3479998

1 INTRODUCTION

The open innovation paradigm appeared formally in 2003 in Chesbrough's book "Open Innovation: The New Imperative for Creating and Profiting From Technology" [7], which stated that companies should start to seriously consider the value in ideas from outside the organization, as well as opening new markets for inside-developed knowledge [7].

Open innovation platforms appeared as a consequence of the spreading of the paradigm, when many organizations felt the need for software to support the functional aspects of Chesbrough's idea. Since software systems are so ubiquitous and pervasive these days, we focus specifically on software-based open innovation platforms.

As the software domain of the open innovation platforms started to mature, more common concepts appeared and became used in such platforms. Reference models can capture these concepts by emphasizing the distinction between functionality in elements and data flow among those elements [2]. Key goals of reference models are facilitating systematic software development [2], and supporting the definition of reference architectures [2][16].

This article proposes a reference model for Software-based Outside-in Open Innovation platforms. A careful review of several reports of open innovation platforms implementations showed two broad contexts in this domain: organizations that open their own innovation process, and organizations that act as open innovation intermediaries (such as brokers, agents, and others). The role of an innovator is engaged in any innovation initiative run in any of these two actors.

We expect this reference model to be used by people involved in the design, implementation and deployment of software-based open innovation platforms, and especially supporting initial decisions regarding platform functional aspects.

The reminder of article is structured as follows. Section 2 describes Open Innovation, the main concepts revolving around the paradigm and related work regarding open innovation platforms definition, design and implementation. Section 3 presents and describes the proposed reference model, with references to relevant sources. Section 4 briefly discusses the research approach. Finally, section 5 summarizes and concludes.

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09…$15.00
https://doi.org/10.1145/3479986.3479998

2 BACKGROUND

As the name implies, open innovation means opening innovation initiatives. In 2003, Henry Chesbrough coined the term in his book "Open Innovation: The New Imperative for Creating and Profiting From Technology" defining it as paradigm in which *"valuable ideas can come from inside or outside the company and can go to market from inside or outside the company"* [7]. Chesbrough contrasted this paradigm with the Closed Innovation one, inwardly focused by its very nature, which is characterized by a vertical integration model where internal innovation activities end up in internally developed products or services that are distributed by the company [9]. Through time, the paradigm and its definition have been dynamic elements, subject to changes and improvements. For example, von Hippel considers open innovation from the point of view of open-source software [9]. For von Hippel, the benefits of an open, distributed innovation eventually allows users, whether they are firms or individual consumers, to innovate by developing what they want and to benefit from innovations developed and freely shared by others [30][31]. "Freely shared" in von Hippel open innovation is key and is something that still needs to get its way through companies [9][11][4].

The open innovation paradigm has been subject to some questioning about whether it is really a phenomenon or just a fad [11]. Later, Chesbrough and others advanced the definition stating that open innovation is *"the purposive use of inflows and outflows of knowledge to accelerate innovation in one's own market, and expand the use of internal knowledge in external markets, respectively."* [13]. The common issue of considering research and development activities separately from business models motivated the appearance of the the idea of Open Business Models which promote a linking between technological innovation and business models [10][14]. More recently, the open innovation community is developing innovation in services, which in some cases is the evolution of products becoming service businesses, and for which service platforms are key [10][8].

A survey of 125 large firms (2,840 companies were invited to participate) in Europe and the United States (annual sales in excess of $250 million) revealed that a 78% of the respondents are practicing open innovation in both low-tech and high-tech sectors (though the same survey also reported that the degree of open innovation practicing is higher in high-tech manufacturing and in trade and retails firms) [11]. In the same survey, informal networking and university research grants ranked second and third, respectively, in terms of the reported importance for inbound open innovation (co-creation with customers and consumers was rated as first). Outbound open innovation practices were, on average, rated as less important than inbound practices, with joint venture activities and sale of market-ready product idea to another organization ranked as first and second, respectively [11].

A follow-up study (121 usable responses, 73 were from European firms) aimed to get more insights from the open innovation project level which is where many of the critical decisions about open innovation are made [4]. This study revealed a 78% of the respondents practicing open innovation with more than 50% of those firms reporting that they adopted the strategy more than five years ago. The same study also revealed that universities and public research organizations were involved as partners in 58% of the projects at the problem definition stage and in 60% of the projects at the solution development phase [4].

Embracing open innovation is not always easy. Many times it requires organizations to be open for, and to work towards, a mindset change from all the involved actors. Concretely, this "new" mindset should be based on a collaborative open culture based on trust [12], science with an "intrapreneurial" attitude [25] and organizations' willingness to invest before starting to see results [24].

2.1 Inside-out and outside-in open innovation

An open innovation endeavor is commonly understood as having two processes: inside-out and outside-in. Some authors recognize a third one, the coupled process [19], which is actually the process archetype that recognizes both incorporating and sharing knowledge is crucial for success [19]. As noted by Chesbrough, open innovation is rarely an only-inbound or only-outbound process [4]. Instead, a mix of the two types of open innovation are seen, notwithstanding in some cases one of the processes is more common [4]. Moreover, as Carroll et al. [6] remark, the opening of an innovation process should be understood as a continuum, that is, from fully open to fully closed we can recognize degrees of openness.

Outside-in open innovation refers to the case in which a company's core innovation approach is based in the integration of external knowledge in the organization [19]. On the other hand, inside-out open innovation refers to focusing on the externalization of company's internally generated knowledge to the outside [19].

2.2 Open innovation intermediaries

Whether being outside-in or inside-out open innovation, organizations that try the paradigm will face some challenges regarding how to make use of the "valuable ideas". Being more precise, organizations making use of open innovation will eventually face:

- Searching and selecting external knowledge and knowledge providers.
- Searching and selecting appropriate markets for internal knowledge to be carried out outside.

This is where open innovation intermediaries come to play a role. Chesbrough [14] define innovation intermediaries as companies focused on helping innovators to use external ideas more rapidly or helping inventors to find more markets to use their own ideas. A key aspect relies in the words "rapidly" and "more" as we cannot neglect the innovators' potential abilities to search for external ideas or markets, but we can appreciate the potential of intermediaries as facilitators in the two aforementioned regards. Another possible benefit is that companies could make use of intermediaries as a means for testing external searches before devoting resources to internally adapt companies' areas to perform external searches [14]. In addition, Hossain [21] recognizes the following more specific tasks for open innovation intermediaries:

- Facilitating internal and external technology commercialization.
- Connecting innovation seekers and innovation providers.
- Helping companies in screening external markets.
- Reducing the costs derived from searching.

- In-licensing, co-developing and acquiring external IP or technologies.

A remarkable point in Hossain's work [21] is that he explicitly links the concepts of intermediaries with innovation platforms.

Howells [23] defines an open innovation intermediary as any "organization or body that acts as an agent or broker in any aspect of the innovation process between two or more parties". In addition, the same author recognizes the following activities as part of the intermediaries work: provide information about potential collaborators, brokering a transaction between two or more parties, acting as a mediator or go-between bodies or organizations, and helping find advice, funding and support for the innovation outcomes of such collaborations [23].

Being an innovation intermediary also comes with challenges as Chesbrough notes [14]:

- Balancing the amount of information available for clients to understand what is the problem and potential solution without disclosing all details that could put a company in a risk of loosing its advantage.
- The problem of identity: how to manage the identity of the participants involved with the intermediary and when, if anonymous, reveal them.
- How to concretely show the value of the service to the clients.
- How to create access in a market with, perhaps a lot of, buyers and sellers.
- How to build trust so clients get convinced to work with an intermediary.

According to Chesbrough [14], there are two broad categories of intermediaries: agents and brokers. Agents represent one side of an exchange of IP or technology. Brokers, also known as market makers, aim to bring parties together. Brokers may also take part in the transaction between parties.

NineSigma[1] is an example of an agent intermediary. NineSigma. The company searches for potential partners for a solution seeker. Solution seeker means that there is a problem behind the search. Both Chesbrough and Hossain agree in that a key service in intermediaries is to aid in the problem definition [14][21].

InnovationXchange[2] is an example of broker intermediary. The company aims to help member companies in sharing requirements and tries to match them to technologies and initiatives residing in other member companies [14]. InnovationXchange explicitly recognizes and works with an innovation network.

As expected, intermediaries categories should not be viewed as strict classes as many intermediaries could eventually slightly touch characteristics of one or another category. Moreover, we expect boundaries in the categories to be even more diffuse in the future as the companies expand their innovation services (for example, as of this writing InnovationXchange also offers technology scouting and technology watch as their services).

The specifics of the revenue streams for these companies is outside the scope of this work. However, we can mention that the mixed "flat-fixed" and "success" fee model is commonly used [21].

2.3 Open innovation platforms

The term "Open Innovation Platform" has been widely used in open innovation literature and initiatives. However, there is no just one definition as the term has been used in many industries and with very varied goals. Still, we can argue that when we talk about a platform here we refer to a software-based platform. Moreover, we mostly refer to an online software-based platform. We can also analyze here common and recurrent published platform functional characteristics. Open innovation intermediaries (see section 2.2) are strongly related to open innovation platforms as their work is mostly supported by this kind of platforms.

Implementing an open innovation platform implies opening the organization to consider ideas and knowledge from external sources such as the crowd. Managers sometimes feel somewhat reluctant to deal with the crowd mainly because they don't completely understand what kinds of problems a crowd can face better and also how to manage the process [3]. A good understanding of how the crowd can be used for open innovation is key. In this regard, Boudreau and Lakhani [3] propose four models: crowd contests, collaborative community, crow complementors, and labor market. In the case of a crowded contest, the key idea is a company that identifies a specific problem, offers some kind of prize (perhaps, cash), and broadcasts an invitation to submit solutions [3]. According to the same authors, contests are also good for solving design problems because creativity and subjectivity influence the evaluation of solutions. In a collaborative community, the idea is that a company decides to join forces with a diverse community, where many of the members are customers and users of a product or service, to innovate in general terms. Practice shows that communities work best when members can share information freely so IP is almost impossible [3]. Unlike crowd contests and collaborative communities, crowd complementors provide solutions to many different problems using the concept of core product or technology [3]. This is somewhat similar to what software architecture tries to achieve with components reuse, but considering that the reuse is based on a core product. This implies that there must be some kind of interface that allows complementors to build new products on this core. Finally, crowd labor markets seek to match buyers and sellers of services by employing conventional non-long-term contracting [3]. Boudreau and Lakhani [3] also recognize that many of the aforementioned ideas are quite old, but technology is being a catalyst for the use of crowd-based approaches to innovation. Again, the reader should note that these models cannot be used for strict classification of open innovation business initiatives as many of the endeavors can slightly touch characteristics from other models.

Malone et al. [29] propose a "genome" model for what they call "collective intelligence." They argue crowd-based platforms and their intrinsic models behind can be described by using a set of "genes" that correspond to particular answers to a set of four questions: what, who, why, and how. For instance, one of the genes of the InnoCentive innovation platform is [what: create scientific solutions, who: crowd, why: money, how: contest] (a set of genes build up the genome of the platform) [29].

[1] https://www.ninesigma.com/
[2] https://www.ixc-uk.com/

In general, online open innovation platforms facilitate the crowd to work collaboratively for ideation, voting, discussions, and suggestion [22]. Also, key characteristics of the crowd in terms of innovation such as being loose and decentralized [3] are, as expected, also observed in open innovation platforms [22]. In addition, open innovation platforms are also used by companies to better engage with customers, suppliers, employees, partners, citizens, and regulators for ecosystem development [22][20].

Two previous works addressed the need for a systematic approach for the design [26], and implementation and deployment [26][15] of open innovation platforms. These works define generic, broad-applicable models to guide the process of enacting open innovation platforms. The generic, broad-applicable characteristics of these proposals make it necessary to consider more concrete guidance in each of the phases they describe. In particular, they require guidance in the potential concrete functional features an open innovation platform could eventually provide, which is the aim of our work.

3 OPEN INNOVATION REFERENCE MODEL

Software-based outside-in open innovation platform contexts can be grouped into two broad classes: (1) organizations (including companies) or a state/government pursuing opening an innovation process, or (2) open innovation intermediaries.

Many of the platform functional characteristics intersect both groups, but we find this distinction useful as their respective platforms users take different roles. For example, in the case of (1), users are usually a subset of a community, whether working for a laboratory, academic setting, or just simply being motivated to engage in the platform for other inner reasons. In the case of (2), users are usually organizations (including companies) seeking intermediary services; in this case, intermediaries are also considered as users (in fact, we would expect the platform to be run and managed by an intermediary).

However, an organization might be deploying an open innovation platform while at the same time seeking services from innovation intermediaries. Therefore, the two groups proposed are presented for practical purposes and should not be taken as an exclusive classification. Moreover, the set of functional characteristics should be seen as one big unified library rather than two separate libraries.

3.1 Functional characteristics

The proposed reference model includes several functional characteristics, all of them wih support in the literature; we summarize then in tables that report both functional characteristics and their sources.

Table 1 summarize functional characteristics that could be expected by open innovation platforms adopters (organizations). Table 2 summarize functional characteristics that could be expected by intermediaries.

Table 1: Open innovation platforms functional characteristics commonly expected by organizations opening their innovation process.

Functional characteristic	Reference
Post a challenge or problem.	[6][22][1][5][17]
Submit proposal.	[6][22][1][5]
Screen proposal.	[6][22][1]
Review proposal.	[6][22][1][5]
Sign IP contract.	[6][22]
Manage IP.	[6][22]
See results (of proposal).	[6]
See recommendations.	[27]
Search database.	[27]
Share knowledge.	[22][1]
Share problems and limitations.	[5]
Vote proposal.	[22]
Send inter-user messages.	[27]
See members details.	[5]
See platform news.	[5]
Manage public profile.	[5]

3.2 Actors

Actors describe roles and not kinds of people or organziations, i.e., the same entity can play different (and several) roles at different times. For example, an open innovation adopter may become an open innovation intermediary or even an innovator.

There may be many subtypes of actors. For example, a university could be working towards opening up its innovation processes, and

Table 2: Open innovation platforms functional characteristics commonly expected by innovation intermediaries.

Functional characteristic	Reference
Post a challenge or problem.	[3][21][28][18][17]
Submit a solution.	[3][21][28][18]
Review a solution.	[3][21][28][18]
Manage challenge or problem.	[3][21][28][18]
Contact innovation partner.	[21]
Sign disclosure agreement.	[21]
Sign IP contract.	[28]
Manage IP.	[28]
Manage innovation partners.	[3][21][23][28][18]
Manage innovation network.	[3][21][23][18]
Manage innovation ideas.	[3][21][23][18]
Manage technology.	[21][23]
Manage technology transaction.	[21]
Manage innovation process.	[23]
Manage innovation consortium.	[21]
See advice about funding.	[23]

thus play a role of open innovation adopter. Nevertheless, for the sake of simplicity, we only consider three generic kinds of actor:

- **An innovator:** Any entity playing a role as an innovation partner in an innovation initiative, a people sharing an idea to be screened for inclusion in an innovation initiative, or a people just collaborating in any stage of an open innovation initiative.
- **An open innovation adopter:** Any entity working towards opening its innovation processes, that is, embracing the open innovation paradigm. Open innovation adopters are expected, most of the time, to be organizations.
- **An open innovation intermediary:** Any entity aiding innovators in outside-in, coupled, or inside-out innovation initiatives. Open innovation intermediaries also work helping innovators in defining their ideas.

3.3 Scenarios

We describe the functional characteristics using a scenario-based approach. Scenario-based approaches are commonly used in software engineering to describe actors making use of a software system in order to achieve some goal.

3.3.1 Post a challenge or problem. Organizations opening their innovation process might expect a platform that allows them to present to the crowd a challenge or problem so the community provide potential proposals for facing or solving the posted challenge or problem. The main and primary actor here is the organization opening its innovation process.

3.3.2 Submit proposal. Organizations that post a challenge or problem for the crowd to engage in its solution will expect the community members to submit proposals which describe potential solutions to the challenge or problem presented. The main and primary actor here is a member of the crowd. More specifically, a peer.

3.3.3 Screen proposal. Organizations might expect several proposals, especially if the platform is widely promoted and managed by a well-known company. Some authors report that before an organization starts to review proposals, they run a screening process before [6][22][1]. The organizations will therefore expect functionality that will allow them to screen proposals. In this sense, the screen proposal functionality will allow the organization to list, search, perhaps filter, and mark as accepted, rejected or delayed a proposal. The main and primary actor here is the organization opening its innovation process.

3.3.4 Review proposal. Screening a proposal aims at an initial selection in order to reduce the amount of proposals to be worked on. After a screening, the platform is expected to provide means for supporting the reviewing of a proposal. At least, the organization should be able to select a proposal for reviewing, managing related documentation and managing the defined stages in the reviewing process. The main primary actor here is the organization opening its innovation process.

3.3.5 Sign IP contract. Although the specifics of intellectual property contractual issues are beyond the scope of this description, we remark that the open innovation platform should provide means for signing and storing an intellectual property contract. The main primary actors here are the organization opening its innovation process and a peer from the crowd.

3.3.6 Manage IP. Signing an IP contract by some means provided by the platform implies the organization will also expect some functionality for managing the IP contract. As in other cases, the particular characteristics of this functionality are a matter of concrete implementations. The main primary actor here is the organization opening its innovation process.

3.3.7 See results of proposal review. The platform should allow a peer from the crowd to see the results of the review of his or her proposal (if previously selected in the screening process). The specifics of the way the organization presents the review results are a matter of concrete platform implementations. The main primary actor here is a peer of the crowd.

3.3.8 See recommendations. The users of the platform should be able to see recommendations regarding some of the many entities worked in the platform. For example, the user could see recommendations of funding, innovation partners, research partners, among other entities. As noted in [27], a platform can provide fast (short) or full recommendations. The main primary actor here is an innovation partner.

3.3.9 Search catalog. The users should be able to search for the entities allowed to be searched in the catalog (e.g., funding applications, projects ideas, among others). Of course, the specific searchable entities is a matter of specific implementations. The main primary actor here is an innovation partner.

3.3.10 Share knowledge. Users of the platform might be able to share knowledge. Here, knowledge refers to any kind of concrete way in which a user could express a belief (e.g., comments about proposals, comments about other topics, opinions, answers to open questions, etc.) Here the main primary actor is an innovation partner. Eventually, a secondary actor could be a moderator.

3.3.11 Share problems and limitations. In some cases such as in [5], open innovation platforms could implement specific ways in which real users with real problems and experiencing limitations to share their thoughts on real problems and limitations. The main primary actor here is an innovation partner.

3.3.12 Vote proposal. If allowed, an open innovation platform could allow users to vote on proposals. The specifics of the voting system and the way the organization will use the votes are a matter of specific implementations. Here the main primary actor is an innovation partner.

3.3.13 Send inter-user message. Organizations might allow users to communicate between them. This functionality should allow users to send messages between them. The main primary actor here is an innovation partner.

3.3.14 See member details. If users register for having accounts and profiles in the platform, the software should allow them to see details about the members. What and how many details should be exposed to members is a matter of concrete implementation. The main primary actor here is an innovation partner.

3.3.15 See platform news. This functionality provides the users with news that the organization would want to share with the community. News can be personalized, private or public. This functional characteristic also implies the organization will be able to store and publish news. Here the main primary actor is an innovation partner.

3.3.16 Manage public profile. If users are required or motivated to register in the platform before making use of other functionality, a public profile is naturally expected. This functional characteristic recognizes a user managing his or her public profile (e.g., adding, deleting or modifying personal info).

3.3.17 Post a challenge or problem. Unlike in the case in which an organization is pursuing opening its innovation process, in this case it is an open innovation intermediary which, after careful working on a submitted problem by a client, can post a challenge or a problem for the innovation network it maintains see it. Main primary actor here is the innovation intermediary which manages and runs the platform.

3.3.18 Submit a solution. If an innovation partner finds a challenger or problem worth to be engaged in, it can submit a solution. Main primary actor here is the innovation partner.

3.3.19 Review a solution. The intermediary should have available functional characteristics that facilitate the review of a submitted solution (by an innovation partner). Main primary actor here is the innovation intermediary.

3.3.20 Manage challenge or problem. A challenge or a problem is expected to be characterized by several attributes. The intermediary should be able to manage the challenge or problem in its attributes or even delete it. Thus, the platform should allow the innovation intermediary to modify or even add attributes and information. Main primary actor here is the innovation intermediary.

3.3.21 Contact innovation partner. Depending on the rules established by the open innovation intermediary, the platform could allow innovation partners to contact other innovation partners. Again, depending on the rules, the contact may be anonymous. Main primary actor here is the innovation partner.

3.3.22 Sign disclosure agreement. Customers will expect the platform provides some means for signing, storing and managing a disclosure agreement. The particular degree of disclosure, whether being complete non-disclosure or disclosure at some degree, is a matter of concrete implementation and even of a particular innovation endeavor. The main primary actor here is the innovation partner. Eventually, the intermediary could be a secondary actor.

3.3.23 Sign IP contract. As in the case in which an organization is opening its innovation process (see scenario 3.3.5), here we consider the functionality to facilitate signing and storing of intellectual contracts. In this case, the main primary actors are the innovation partner and the intermediary.

3.3.24 Manage IP. As in the case in which an organization is opening its innovation process (see scenario 3.3.6), the intermediary will expect some means for managing the IP contract. Unlike the aforementioned case, here the main primary actor is the open innovation intermediary.

3.3.25 Manage innovation partner. This functionality treats the innovation partner as an important entity in the platform. As the open innovation intermediary will deal with many partners (in more general terms, clients), the platform is expected to provide some means for managing individual innovation partners. In addition, innovation partners could be also part of innovation networks also managed by the intermediary. The main primary actor is the intermediary.

3.3.26 Manage innovation ideas. Innovation ideas are also treated as entities in the platform. This functionality reveals generic characteristics related to managing innovation ideas: reception (if sent for evaluation by an evaluation partner), modification, assignation, among others. The main primary actor here is the intermediary.

3.3.27 Manage technology. As in the case of innovation ideas, this functionality also treats technology as an entity subject to management in the platform. The main primary actor here is the intermediary.

3.3.28 Manage technology transaction. As open innovation intermediaries could also offer technology brokering services, an open innovation platform could also provide means for managing technology transactions. Transactions can take the form of processes which should also be considered when providing functionality for managing technology transactions. Main primary actor here is the intermediary.

3.3.29 Manage innovation process. This functional characteristic proposes typical characteristics for managing an innovation process in which many innovation partners could engage. For example, managing activities, roles, milestones, work products, among other entities. The main primary actors here are the intermediary and the innovation partners.

3.3.30 Manage innovation consortium. Whenever the intermediary forms a consortium, the platform should provide functionality for collaborative working between the innovation partners. Document sharing, innovation partners incorporation, among others are examples of expected characteristics. The main primary actor here is the intermediary.

3.3.31 See advice about funding. In case in which the intermediary running the platform would like to share funding applications available, the platform should provide functionality for an innovation partner to see this information. The particular characteristics to be considered are a matter of concrete implementations (e.g., filtering, searching). The main primary actor here is the innovation partner.

4 RESEARCH APPROACH

In the context of a software-based open innovation platform implementation, we deemed necessary to count with guidance about expected functional characteristics for the software system supporting these kind of platforms.

We started searching for peer-reviewed papers, book chapters, and articles published in well-known sources such as Harvard Business Review.

For the peer-reviewed papers we used the following sources: Springer[3], ScienceDirect[4], ACM Digital Library[5] and IEEE Xplore[6] . In addition, Google Scholar[7] indexes journals (e.g. Research-Technology Management by Taylor & Francis) that proved to be valuable resources.

Google search engine was used to find other non-peer-reviewed articles and book chapters. We only deemed to be acceptable for this paper articles that appeared published in well-known, reputable sources (e.g. Harvard Business Review).

The articles and book chapters were first screened by one of the authors. In this screening process, some papers were removed mainly due to lack of validation and consistency problems. Citations, as noted by Google Scholar, was also used as an indicator for removing or accepting a paper in this first screening process.

In a second round, the articles and book chapters that reached this phase were read and analyzed. The analysis allowed us to build this proposed reference model. This phase was carried out by all of the authors.

A third phase consisted in a cross-checked review from all of the authors. This third phase also ended up in improving the structure and argument of the reference model.

5 CONCLUSIONS AND FUTURE WORK

There is enough support in literature to argue that open innovation is a paradigm that is more than a simple fad. It has proven to be very useful for many organizations, especially large companies, mainly due to the access to externally created knowledge, an asset that is not always found within the organization. This search for talent and collaboration is necessary today, in a market in which aspects such as reducing costs, risks, and deadlines, in addition to increasing competitive advantage goals, are essential for the companies to survive. Innovating instead of reducing costs more than the competition, seems to be one of the formulas to survive in this changing market in which we live.

We hope that this model can serve as a starting point and as a guide for organizations that are committed to implementing and deploying a software-based outside-in open innovation platform. We expect the leader for these initiatives find useful information in this reference model which can be also associated to other proposals already published to aid in the journey to having one of these platforms in production.

As this reference model clearly resembles the main open innovation related tasks, we believe this proposal could also greatly contribute to understanding innovation processes, as well as identifying and regulating relationships between different and multiple stakeholders involved in these processes.

Having defined this reference model, we plan, as a next step, to use this proposal for defining a reference architecture for open innovation platforms.

ACKNOWLEDGMENTS

This work was partially funded by ANID PIA/APOYO AFB180002 and by UTFSM under DPP grants.

REFERENCES

[1] Sabrina Adamczyk, Angelika C. Bullinger, and Kathrin M. Moeslein. 2011. Commenting for new ideas: insights from an open innovation platform. *International Journal of Technology Intelligence and Planning* 7, 3 (Jan. 2011), 232–249. https://doi.org/10.1504/IJTIP.2011.044612

[2] Samuil Angelov, Paul Grefen, and Danny Greefhorst. 2012. A framework for analysis and design of software reference architectures. *Information and Software Technology* 54, 4 (2012), 417 – 431. https://doi.org/10.1016/j.infsof.2011.11.009

[3] Kevin J. Boudreau and Karim R. Lakhani. 2013. Using the Crowd as an Innovation Partner. *Harvard Business Review* April 2013 (April 2013). https://hbr.org/2013/04/using-the-crowd-as-an-innovation-partner

[4] Sabine Brunswicker and Henry Chesbrough. 2018. The Adoption of Open Innovation in Large Firms. *Research-Technology Management* 61, 1 (2018), 35–45. https://doi.org/10.1080/08956308.2018.1399022

[5] Angelika C. Bullinger, Matthias Rass, Sabrina Adamczyk, Kathrin M. Moeslein, and Stefan Sohn. 2012. Open innovation in health care: Analysis of an open health platform. *Health Policy* 105, 2 (2012), 165 – 175. https://doi.org/10.1016/j.healthpol.2012.02.009

[6] Glenn P. Carroll, Sanjay Srivastava, Adam S. Volini, Marta M. Piñeiro-Núñez, and Tatiana Vetman. 2017. Measuring the effectiveness and impact of an open innovation platform. *Drug Discovery Today* 22, 5 (2017), 776 – 785. https://doi.org/10.1016/j.drudis.2017.01.009

[7] Henry Chesbrough. 2003. *Open Innovation: The New Imperative for Creating and Profiting From Technology.* Harvard Business School Press.

[8] H. Chesbrough. 2010. *Open Services Innovation: Rethinking Your Business to Grow and Compete in a New Era.* Wiley.

[9] Henry Chesbrough. 2012. Open Innovation: Where We've Been and Where We're Going. *Research-Technology Management* 55, 4 (2012), 20–27. https://doi.org/10.5437/08956308X5504085

[10] Henry Chesbrough. 2017. The Future of Open Innovation. *Research-Technology Management* 60, 1 (2017), 35–38. https://doi.org/10.1080/08956308.2017.1255054

[11] Henry Chesbrough and Sabine Brunswicker. 2014. A Fad or a Phenomenon?: The Adoption of Open Innovation Practices in Large Firms. *Research-Technology Management* 57, 2 (2014), 16–25. https://doi.org/10.5437/08956308X5702196

[12] Henry Chesbrough, Sohyeong Kim, and Alice Agogino. 2014. Chez Panisse: Building an Open Innovation Ecosystem. *California Management Review* 56, 4 (2014), 144–171. https://doi.org/10.1525/cmr.2014.56.4.144

[13] H. Chesbrough, W. Vanhaverbeke, and J. West. 2008. *Open Innovation: Researching a New Paradigm.* OUP Oxford. https://books.google.cl/books?id=lgZAyauTEKUC

[14] H.W. Chesbrough and H. W. 2006. *Open Business Models: How to Thrive in the New Innovation Landscape.* Harvard Business School Press. https://books.google.cl/books?id=FzWqNyPtC38C

[15] Pablo Cruz and Hernán Astudillo. 2020. Towards a Maturity Model for Assessment of Organization Readiness in Implementing and Deploying an Open Innovation Platform. In *Proceedings of the 16th International Symposium on Open Collaboration* (Virtual conference, Spain) *(OpenSym 2020)*. Association for Computing Machinery, New York, NY, USA, Article 12, 4 pages. https://doi.org/10.1145/3412569.3412868

[16] Lucas Bueno Ruas de Oliveira, Katia Romero Felizardo, Daniel Feitosa, and Elisa Yumi Nakagawa. 2010. Reference Models and Reference Architectures Based on Service-Oriented Architecture: A Systematic Review. In *Software Architecture*, Muhammad Ali Babar and Ian Gorton (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 360–367.

[17] Enrique Estellés-Arolas and Fernando González-Ladrón de Guevara. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science* 38, 2 (2012), 189–200. https://doi.org/10.1177/0165551512437638

[18] Karsten Frey, Christian Lüthje, and Simon Haag. 2011. Whom Should Firms Attract to Open Innovation Platforms? The Role of Knowledge Diversity and Motivation. *Long Range Planning* 44, 5 (2011), 397 – 420. https://doi.org/10.1016/j.lrp.2011.09.006 Social Software: Strategy, Technology, and Community.

[19] Oliver Gassmann and Ellen Enkel. 2004. Towards a Theory of Open Innovation: Three Core Process Archetypes. In *R&D Management Conference (RADMA 2004)*.

[20] Francis J. Gouillart and Douglas W. Billings. 2013. Community-powered problem solving. *Harvard Business Review* 91 4 (2013), 70–7, 140.

[21] Mokter Hossain. 2012. Performance and Potential of Open Innovation Intermediaries. *Procedia - Social and Behavioral Sciences* 58 (2012), 754 – 764. https://doi.org/10.1016/j.sbspro.2012.09.1053 8th International Strategic Management Conference.

[22] Mokter Hossain and K. M. Zahidul Islam. 2015. Ideation through Online Open Innovation Platform: Dell IdeaStorm. *Journal of the Knowledge Economy* 6, 3

[3] Springer: https://www.springer.com/

[4] ScienceDirect: https://www.sciencedirect.com/

[5] ACM Library: https://dl.acm.org/

[6] IEEE Xplore: https://ieeexplore.ieee.org

[7] Google Scholar: https://scholar.google.com/

(Sept. 2015), 611–624. https://doi.org/10.1007/s13132-015-0262-7

[23] Jeremy Howells. 2006. Intermediation and the role of intermediaries in innovation. *Research Policy* 35, 5 (2006), 715 – 728. https://doi.org/10.1016/j.respol.2006.03.005

[24] Larry Huston and Nabil Sakkab. 2007. Implementing Open Innovation. *Research-Technology Management* 50, 2 (2007), 21–25. https://doi.org/10.1080/08956308.2007.11657426

[25] Robert Kirschbaum. 2005. Open Innovation In Practice. *Research-Technology Management* 48, 4 (2005), 24–28. https://doi.org/10.1080/08956308.2005.11657321

[26] Roberto Osorno and Norma Medrano. 2020. Open Innovation Platforms: A Conceptual Design Framework. *IEEE Transactions on Engineering Management* (2020), 1–13. https://doi.org/10.1109/TEM.2020.2973227

[27] J. Protasiewicz. 2017. Inventorum: A platform for open innovation. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 10–15.

[28] Jan Henrik Sieg, Martin W. Wallin, and Georg Von Krogh. 2010. Managerial challenges in open innovation: a study of innovation intermediation in the chemical industry. *R&D Management* 40, 3 (2010), 281–291. https://doi.org/10.1111/j.1467-9310.2010.00596.x

[29] Chrysanthos Dellarocas Thomas W. Malone, Robert Laubacher. 2010. The Collective Intelligence Genome. *MIT Sloan Management Review* (2010). https://sloanreview.mit.edu/article/the-collective-intelligence-genome/

[30] Eric von Hippel. 2005. Democratizing innovation: The evolving phenomenon of user innovation. *Journal für Betriebswirtschaft* 55, 1 (March 2005), 63–78. https://doi.org/10.1007/s11301-004-0002-8

[31] E. von Hippel and MIT. Press. 2005. *Democratizing Innovation*. MIT Press. https://books.google.cl/books?id=BvCvxqxYAuAC

Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia

Luís Couto
Faculty of Engineering of the University of Porto
Porto, Portugal
mieic1204994@fe.up.pt

Carla Teixeira Lopes
INESC TEC, Faculty of Engineering of the University of Porto
Porto, Portugal
ctl@fe.up.pt

ABSTRACT

Wikipedia is a free, multilingual, and collaborative online encyclopedia. Nowadays, it is one of the largest sources of online knowledge, often appearing at the top of the results of the major search engines, being one of the most sought-after resources by the public searching for health information. The collaborative nature of Wikipedia raises security concerns since this information is used for decision-making, especially in the health area. Despite being available in hundreds of idioms, there are asymmetries between idioms, namely regarding their quality. In this work, we compare the quality of health information on Wikipedia between idioms with 100 million native speakers or more, and also in Greek, Italian, Korean, Turkish, Persian, Catalan and Hebrew, for historical tradition. Quality metrics are applied to health and medical articles in English, maintained by WikiProject Medicine, and their versions in the above idioms. With this, we contribute to a clarification of the role of Wikipedia in the access to health information. We demonstrate differences in both the quantity and quality of information available between idioms. English is the idiom with the highest quality in general. Urdu, Greek, Indonesian, and Hindi achieved lower values of quality.

CCS CONCEPTS

- **Applied computing** → **Consumer health**; Health informatics;
- **Information systems** → **Wikis**.

KEYWORDS

Information Quality, Wikipedia, Health information, Multilingual information access

ACM Reference Format:
Luís Couto and Carla Teixeira Lopes. 2021. Equal opportunities in the access to quality online health information? A multi-lingual study on Wikipedia. In *17th International Symposium on Open Collaboration (OpenSym 2021), September 15–17, 2021, Online, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3479986.3480000

1 INTRODUCTION

Online health information-seeking behavior improves the patient-physician relationship and patients' engagement in health decision-making [32]. Three health-related terms are in the top ten Google searches [13] for 2020 - "coronavirus", "coronavirus update" and "coronavirus symptoms". A study conducted by the Health On the Net Foundation [27] shows that when information needs relate to health, 44% of respondents admitted looking for this information more than three times a week, with the main point of access being search engines, which may end up leading them later to Wikipedia [18]. The same study revealed that the quality of information remains the most significant barrier encountered by respondents (80%) when searching for health information online, and the factor most valued regarding the quality of information is the reliability/credibility (96%). Different authors have concluded that Wikipedia is a reliable source of health information in surgical information [6], pediatric otolaryngology [35], pharmacology [17] and cancer [28].

Nowadays, the most popular health article - "COVID-19 pandemic", has, on average, more than 40 thousand daily views [43]. Wikipedia was created in English, and the second idiom after that was German, followed immediately by Catalan, these remaining the only idioms for two months. By the end of the first year, Wikipedia had articles written in 18 different idioms. It is currently available in 321 idioms, with 310 of them active [39].

Given existing differences in access to health information among speakers of different idioms [1], Wikipedia can potentially reduce or accentuate this imbalance. In this work, we will compare the quality of information available to users speaking different idioms, checking if it is similar between the different versions or whether there are disparities, and if so, to be able to quantify them.

Section 2 describes how quality is assured in an open, collaborative resource such as Wikipedia. In Section 3, we describe differences between idioms at Wikipedia. Our methodology to compare the quality of Wikipedia health contents in several idioms is described in Section 4. Results and respective discussion are presented in Sections 5 and 6, respectively. Finally, conclusions are presented in Section 7.

2 WIKIPEDIA INFORMATION QUALITY

Quality has always been a concern for Wikipedia, which has established frameworks to ensure it ever since its creation. Wikipedia currently has more than 400 million articles, so assessing the quality of so much information asks for automation. Several authors have addressed this issue, one of the most prominent being Stvilia *et al.* [30].

OpenSym 2021, September 15–17, 2021, Online, Spain

ACM ISBN 978-1-4503-8500-8/21/09.
https://doi.org/10.1145/3479986.3480000

2.1 Wikipedia internal quality mechanisms

Despite the large number of articles created initially, they did not have the desired quality, which led Larry Sanger to define rules published on the Wikipedia pages "Wikipedia is not a dictionary"[1] and "What Wikipedia is not"[2], which still exist, with changes over time. In this context, five principles define the rules and recommendations for preparing content [42]. The first principle states that "Wikipedia is an encyclopedia", pointing that it combines features of encyclopedias, almanacs, and gazetteers; the second principle refers that "Wikipedia is written from a neutral point of view", indicating that articles should have an impartial tone, documenting and explaining significant points of view; the following principle declares that "Wikipedia is free content that anyone can use, edit, and distribute", evidencing that authors freely license their work to the public; the fourth principle expresses that "Wikipedia's editors should treat each other with respect and civility", denoting the etiquette all users should use; the last principle states that "Wikipedia has no firm rules", signifying that Wikipedia policies and guidelines are flexible and mutable over time.

These five pillars are common to Wikipedias in different idioms, but policies are defined for each version. These policies are created by the community, by consensus or by vote, with a transversal character to all the articles present and all its users. There are sanctions for those who violate them, such as blocking users for some time [40].

There are control mechanisms to ensure compliance with these principles that can be summarized into nine types. First, there are many users, where the well-meaning vastly outnumber the malicious, with their unique characteristics working together for a typical result. It is the supervision of users. Next, many editors guarantee neutrality and different points of view on the one hand and, on the other hand, ease in repairing errors. This is the collaborative knowledge construction mechanism. The next control mechanism relates to the fact that there is only one page for everyone, pressuring for a consensus among all and the desired neutrality. Also noteworthy is that no superior entities control the content, avoiding manipulations motivated by secondary interests. It is the wiki structure. Another control mechanism relates to the rules, policies, and principles, defined to ensure good conduct on the one hand and ensure on the other the disruptive potential necessary for evolution. It is the respect for policies and principles. For the next control mechanism, we refer to the concerns and opinions of minorities, taken into account in trying to reach a decision that reflects the values of the community. It is the consensus-based *ethos*. There are intrinsic escalation mechanisms, such as that users will more closely watch items that are more prone to vandalism to stop it. There are also extrinsic mechanisms, such as the possibility that anyone can request disputes in progressive stages. This control mechanism is the escalation and dispute resolution processes. The next mechanism refers to the software tools used by the most active editors, such as Huggle[3], to automatically detect vandalism in real-time, among other tools facilitating the identification and correction of quality problems. It is the software facilitating monitoring and editing control mechanism. Tools exist to block problematic publishers and protect pages from low-quality publishers, capable of filtering combinations of accounts or IP addresses. This control mechanism refers to blocking and protection systems. Finally, inline tags can be used in the text to individual flag statements, individual statements, quotes, or articles as a whole, request verification or citation, and indicate to other users that a fact or presentation is not supported as is. It is the categorization of information control mechanism.

The various versions of Wikipedia generally have an article quality assessment system [41] that is not standardized. For example, in the English version, this system is based on letters that indicate how complete an article is, taking into account different factors. WikiProjects[4] members assess quality using tags that can be used to generate statistical data about the articles. These assessments make it possible to determine the quality of the information in specific areas and prioritize articles for improvement according to expectations. It should be noted that this evaluation has no official character. In addition, there may also be a ranking of the priority or importance of an article, reflecting the level of expectation or desire that a particular topic is portrayed. The scale generally ranges from "unimportant" to "extremely important". This importance rating is also relative to each WikiProject.

2.2 Metrics for assessing quality

To assess Wikipedia information quality, authors propose different metrics based on different features. Generically, Wu *et al.* [44] used four groups of metrics, with a total of 28 metrics: lingual - e.g., readability; structural - e.g., links; historical - e.g., article age and reputational - e.g., amount of editors. Li *et al.* [20] and De La Robertie *et al.* [5] proposed metrics based on the relationship between articles and their editors. Marrese-Taylor *et al.* [22], in 2019, based their work on the articles' editions, also considering the description of each edition.

In the health area, there have been other approaches by authors such as Thomas [33], in 2013, using: comprehensibility - the ratio of medical codes in articles; trust - number of references in articles and readability. In 2014, Conti *et al.* [3] assessed 2,400 medical articles using metrics from types: lingual - Flesch Reading Ease and Flesch-Kincaid scales; structural - e.g., number of links and citations; historical - e.g., number of editions and number of editors; reputational: e.g., age of editors and duration of editions. Modiri *et al.* [25] assessed articles in the neurosurgery area, using readability indexes, the "Center for Disease Control Clear Communication Index"[5] and DISCERN[6]. Later, in 2019, Suwannakhan *et al.* [31] assessed the information quality in anatomy, with readability indexes and DISCERN in association with Wikimedia X-tools [7]. In the same year, Domingues and Teixeira Lopes [7] compared the quality of the Portuguese version with the Anglophone version of Wikipedia in articles related to medicine. They used metrics defined by Stvilia *et al.* [30] and more specific metrics, such as the number of medicine templates, number of medicine infoboxes, and number of citations. Later, in 2021, Couto and Teixeira Lopes [4]

[1] https://en.wikipedia.org/wiki/Wikipedia:Wikipedia_is_not_a_dictionary
[2] https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not
[3] https://en.wikipedia.org/wiki/Wikipedia:Huggle
[4] https://en.wikipedia.org/wiki/Wikipedia:WikiProject
[5] https://www.cdc.gov/ccindex/index.html
[6] http://www.discern.org.uk/discern_instrument.php
[7] https://www.mediawiki.org/wiki/XTools

evaluated the quality of health-related Wikipedia articles, using the same metrics from Stvilia, with the proposal to add health-related features of Wikipedia articles, such as health templates, medical codes, or recommended sections.

Stvilia *et al.* [30] defined seven metrics: authority, completeness, complexity, informativeness, consistency, currency, and volatility to assess the Wikipedia quality. These metrics use 19 features from Wikipedia articles and their history. We consider the metrics and their features, as defined by the authors, complete and comprehensive, as they include several dimensions of the quality of information on Wikipedia. They are also, specific in the form of calculation. They are, therefore, a reference for other authors in different works [2, 7, 15, 19, 44]. Considering this, we use the same metrics and respective features as described in Section 5.

3 DIFFERENCES BETWEEN IDIOMS

There are currently 7,139 living idioms worldwide [9]. Such diversity can naturally raise questions about their presence on the web. In 2009, Pimienta *et al.* [26] described an investigation carried out from 1996 to 2008 by UNESCO through FUNREDES and Union Latine on linguistic diversity on the web that revealed a large discrepancy in the presence of idioms in cyberspace.

Based on language as the primary mean of communication, Wikipedia is an indicator of online multilingualism across the range of idioms present. The Wikimedia Foundation has defined policies for proposing new idioms, created by its "Language committee" [37], responsible for processing the proposals and associated projects. Proposed idioms must not yet exist on Wikimedia, must have a valid ISO 639-1 code, and must have a sufficient number of fluent users to form a viable community of contributors and audiences interested in its content. Regional dialects and different forms of it are excluded. For approval, it is also a requirement that a test project exists on Wikimedia and that there is an ongoing effort to translate the Wikimedia interface into that idiom. There are currently 273 requests for new idioms to be added to Wikipedia [38].

3.1 Wikipedia articles and users

Using the statistical data provided by Wikipedia [39] as of May 2021, Table 1 was adapted, showing the number of articles, edits to articles, administrators, and active users for each version of Wikipedia in idioms with more than 1 million articles.

We can conclude that there are currently 18 idioms on Wikipedia with more than 1 million articles. The superiority of the English idiom in terms of the number of articles available is quickly confirmed. Cebuano, an Austronesian idiom spoken in the Philippines by 27.5 million people in 2020, follows. Cebuano's popularity happens because a bot, Lsjbot, created more than 17 million articles, accounting for most of the articles written in Cebuano, Swedish, and Waray, which also explains why Swedish comes third, followed only then by German. This bot activity also explains why the number of articles does not keep up with the other metrics in the table, where English also stands out, followed by German and French.

In 2009, Dijk [34] addressed Wikipedia edits in minority idioms and ways to measure them for comparison. He mentions the obsession with the number of articles in each Wikipedia and the comparison with other idioms. He concluded that it is difficult to attribute the factors that contribute to the growth of each version of Wikipedia but emphasizes the number of speakers, as they represent the potential article editors of that idiom. This, however, not always corresponds to reality [17]. In 2017, Matei [24], using data from edits from the first decade of Wikipedia's existence, concluded that only 1% of the editors created 77% of the articles, which raises problems about its collaborative spirit. Dijk mentions the importance of people's attitude towards projects such as Wikipedia, pointing this as the main factor for the growth of Latin idioms in Wikipedia. He concludes with the importance of the collaboration of institutions related to idiom issues in content development, especially in minority idioms. Later in 2011, Hale [14] studied the role of multilingual editors as enablers of the development of the various idioms within Wikipedia.

Table 1: Wikipedia statistics for idioms with more than 1 million articles

	Articles	Edits	Admins	Active users
English	6,296,349	1,018,157,853	1,096	138,226
Cebuano	5,729,196	31,370,481	6	170
Swedish	3,187,113	49,166,960	63	2,617
German	2,575,270	210,442,253	187	20,382
French	2,327,445	182,377,650	156	21,964
Dutch	2,054,789	58,767,964	35	4,343
Russian	1,723,112	113,865,981	79	11,723
Italian	1,692,357	120,298,053	114	9,911
Spanish	1,682,915	135,034,634	67	17,133
Polish	1,473,158	63,082,287	102	4,802
Egyptian Arabic	1,283,253	5,590,058	6	210
Japanese	1,267,954	83,259,311	41	15,260
Waray	1,265,315	6,233,530	3	76
Vietnamese	1,263,818	64,842,686	20	2,300
Chinese	1,196,344	65,280,127	79	8,365
Arabic	1,115,708	53,691,546	27	5,422
Ukrainian	1,091,668	31,568,224	45	3,442
Portuguese	1,066,210	60,993,351	71	10,358

Source: adapted from https://en.wikipedia.org/wiki/List_of_Wikipedias

The Wikidata[8] inter-language system is a system launched in 2012 by Wikimedia Foundation, which together with the inter-linguistic links[9] provides a centralized solution based on a collaborative database. It allows connecting the same concept across multiple versions of Wikipedia and even between other Wikimedia projects. Essentially, items are stored, each with a label, a description, and a list of alternative names, linking the items and their data together. Hale found that most editors are active in only one idiom, with 15% doing so in different idioms, and they are usually more active than others.

3.2 Content quality

In 2009, Filatova [10] described the multilingualism of Wikipedia through a framework created for this purpose and using only the text of the articles. The author mentions that articles about the same thing differ a lot between versions, especially in terms of the amount of information covered in each version and the aspects that authors choose to cover about the general topic of the article, directly affecting its quality. Domingues and Teixeira Lopes [7] conducted a comparative study on the quality of medicine-related

[8] https://www.wikidata.org/wiki/Wikidata:Wikidata_Concepts_Monitor
[9] https://en.wikipedia.org/wiki/Help:Interlanguage_links

articles in the Portuguese and English versions of Wikipedia in 2019. The authors found significant differences between the two versions in the vast majority of the metrics evaluated. The results suggest that English articles demonstrate more significant effort in content organization, information reuse, and citation usage. The overall conclusion is that Wikipedia's English health contents are substantially better in terms of quality.

Despite the scarce research available on the differences in content quality between different Wikipedia versions according to idioms, there seems to be a direct relationship between the quantitative and qualitative aspects of the information available. Assuming that idioms with lower quantitative expression in Wikipedia translate lower quality information, and given the importance of Wikipedia as a source of information, this is an inequality problem that has received the attention of UNESCO, which recognizes that the information present in cyberspace is a significant factor for the development of humanity, as it is a primary way of sharing information and knowledge.

4 METHODOLOGY

Our approach has five major steps, schematized in Figure 1. Numbers identify the execution sequence, and arrows identify information flow.

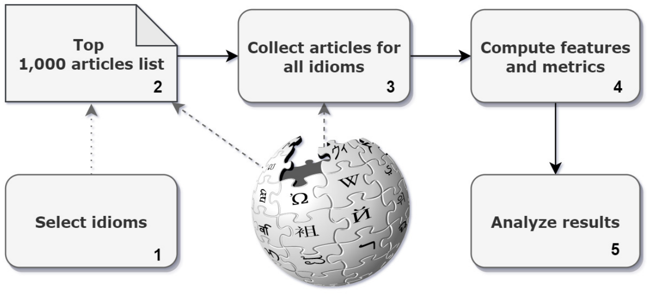


Figure 1: Methodology

We began by selecting the idioms for our dataset. Then, we collected a list of health-related articles. These first steps are described in Subsection 4.1. In the next step, we gathered the articles in that list for all idioms, as described in Subsection 4.2. After that, we assessed the quality of the articles, as described in Section 5 and finally, we discussed the results, as described in Section 6.

4.1 Idioms selection

We have selected idioms available on Wikipedia with at least 100 million speakers as a native or second idiom. We also extended this collection to six other idioms for their cultural or medical importance, namely their neurosurgical tradition since ancient times: Greek, Italian, Korean, Turkish, Persian, and Hebrew.

In Table 2, we can see the number of speakers for each idiom of our dataset, as first or second idiom, sorted by decreasing number of total speakers. English is the idiom with most speakers, mostly as a second idiom. Chinese follows closely, mostly as first idiom. In third place, Hindi comes with a significant difference from the previous two. Hebrew and Catalan are the less spoken idioms in our dataset, with only 9 million total speakers.

Table 2: Number of speakers for each idiom of our dataset

	First idiom	Second idiom	Total
English (en)	369.9 million	978.2 million	1.348 billion
Chinese (zh)	921.2 million	198.7 million	1.120 billion
Hindi (hi)	342.2 million	258.3 million	600 million
Arabic (ar)	-	-	274 million
Bengali (bn)	228.7 million	39.0 million	268 million
French (fr)	79.6 million	187.4 million	267 million
Russian (ru)	153.7 million	104.3 million	258 million
Portuguese (pt)	232.4 million	25.2 million	258 million
Urdu (ur)	69.0 million	161.0 million	230 million
Indonesian (id)	43.6 million	155.4 million	199 million
German (de)	76.6 million	58.5 million	135 million
Japanese (ja)	126.3 million	121,500	126 million
Turkish (tr)	82.2 million	5.9 million	88 million
Persian (fa)	56.3 million	17.9 million	74 million
Korean (ko)	-	-	82 million
Italian (it)	64.8 million	3.1 million	68 million
Greek (el)	-	-	13 million
Hebrew (he)	-	-	9 million
Catalan (ca)	-	-	9 million

4.2 Data collection

Our selection of health-related articles was based on a list maintained by WikiProject Medicine [43]. This list contains the 1,000 most viewed articles for the English Wikipedia.

First, all articles written in English were collected from the mentioned list. Data for articles written in other idioms other than English was obtained by following the idiom link in each of the English articles, and each of them was iteratively collected.

We used the MediaWiki API to collect the article's contents and metadata, revision history, idiom links, internal links, and external links, following the approach of Domingues and Teixeira Lopes [7]. Other data was obtained from the article's markup. We also obtained images through markup because the API does not distinguish content images from others, such as Media Wiki and Wikimedia logos. Templates, infoboxes, and citations were also collected from the article's markup. To compute some measurements, such as readability scores, "InfoNoise", or the article's length, we removed all the markup from the article's content to obtain the required plain text. We faced some challenges when collecting data because there is considerable heterogeneity among the idioms chosen. There is also heterogeneity between the different versions of Wikipedia for each idiom. Moreover, there is also heterogeneity within each Wikipedia version, as edits are made by several users, who do not always comply with the established standards when they exist.

As some articles only have versions in some idioms, the complete dataset consists of 14,456 articles. The distribution of the articles by idiom is visible in Figure 2. This figure also includes the distribution of all Wikipedia articles by idiom. Figure 2 shows that English is the only idiom with 1,000 articles, meaning that no other idiom has the corresponding version for all the articles in the top list. It is also evident the prominent differences between idioms, where some only have about half, or even less, of the total number of articles, such as the Urdu idiom. Analyzing the relation of the dataset number of

Table 3: Idioms quality assessment for authority metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	2033.05	1196.7	de	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn	18
German	de	1315.5	416.7	ru	it	zh	fr	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn		17
Russian	ru	1250.8	265.6	zh	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Italian	it	1240.2	254.7	zh*	hi	pt	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn					14
Chinese	zh	1230.4	344.5	hi	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn							12
French	fr	1189.8	233.8	hi	pt*	tr	he	ar	ja	ca	ur	fa	id	ko	el	bn						13
Hindi	hi	1159.6	546.1	pt	ur	id	ko*	el	bn													6
Portuguese	pt	1152.6	273.5	ar	ja	ca	ur	fa	id	ko	el	bn										9
Turkish	tr	1148.5	539.6	ur	fa*	id	ko	el	bn													6
Hebrew	he	1139.3	413.5	ur	fa*	id	ko	el	bn													6
Arabic	ar	1130.3	603.0	ur	id	ko	el	bn														5
Japanese	ja	1128.4	186.8	ur	id	ko	el	bn														5
Catalan	ca	1101.6	468.0	ur	id	ko	el	bn														5
Urdu	ur	1096.6	994.0	fa																		1
Persian	fa	1087.2	490.4	id	ko	el	bn															4
Indonesian	id	1034.5	1086.8																			0
Korean	ko	1024.3	956.2	bn*																		1
Greek	el	800.8	1069.9																			0
Bengali	bn	710.1	1070.2																			0
χ^2		3543.3																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the remaining values

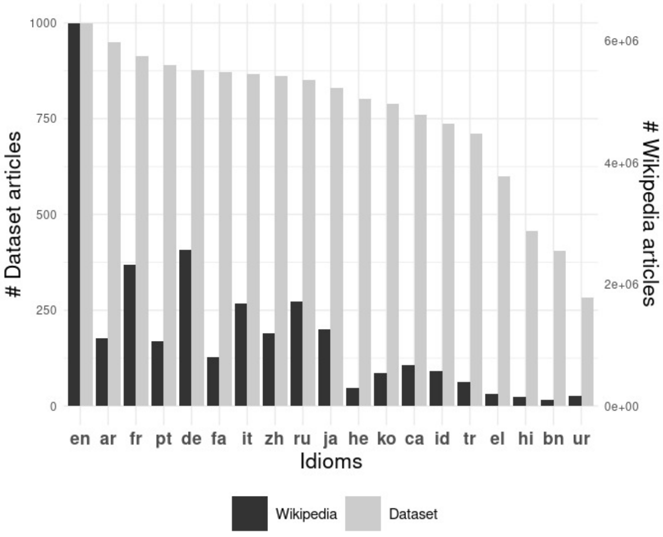


Figure 2: Number of articles by idiom

articles and the total number of articles in each version of Wikipedia, we can observe different distributions. Given that the dataset only contains health-related articles, this data suggests that the size of each version of Wikipedia is not directly related to the number of health-related articles. These data may also point to differences in the importance of the health-related area in each version of Wikipedia.

The datasets and code developed in this work are publicly available in an institutional repository. [10], [11]

4.3 Data analysis

We compared the several idioms in terms of metrics and their features. As most of the metrics and features do not follow a normal distribution in each idiom, we used the median as our measure of central tendency and the interquartile range as our dispersion measure. As the assumptions for the one-way analysis of variance (ANOVA) hypothesis test were not verified, we applied the Kruskal-Wallis to realize if there were significant differences between idioms in each feature and metric. If so, we performed *post-hoc* tests, namely the Dunn pairwise test, with *p*-values adjusted by the Holm method, to identify the significant differences. When reporting our results, we use * to indicate results significant at an alpha=0.05 and ** to indicate results significant at an alpha=0.001.

5 RESULTS

Our results are organized by metric and described in the following sections.

5.1 Authority

Authority is "the degree of the reputation of an information object in a given community" [29], and it is computed as: ***Authority = 0,2 ∗ Num. Unique Editors + 0,2 ∗ Num. Edits + 0,1 ∗ Connectivity + 0,3 ∗ Num. Reverts + 0,2 ∗ Num. External Links + 0,1 ∗ Num. Registered User Edits + 0,2 ∗ Num. Anonymous User Edits***. The number of unique editors corresponds to the number of different authors involved in the article's editions and is extracted from its history. Connectivity corresponds to the number of articles connected to a particular article through common editors and is obtained from each article's editors and the articles edited by them. This metric has the drawback of being based solely on articles in the database, requiring a large dataset to be accurate. Reverts correspond to the number of reversions made to editions of the article, and it is based on its editing history. External links correspond to the number of links in the article that points to content outside

[10] https://doi.org/10.25747/ep0v-en19
[11] https://doi.org/10.25747/wfzk-h937

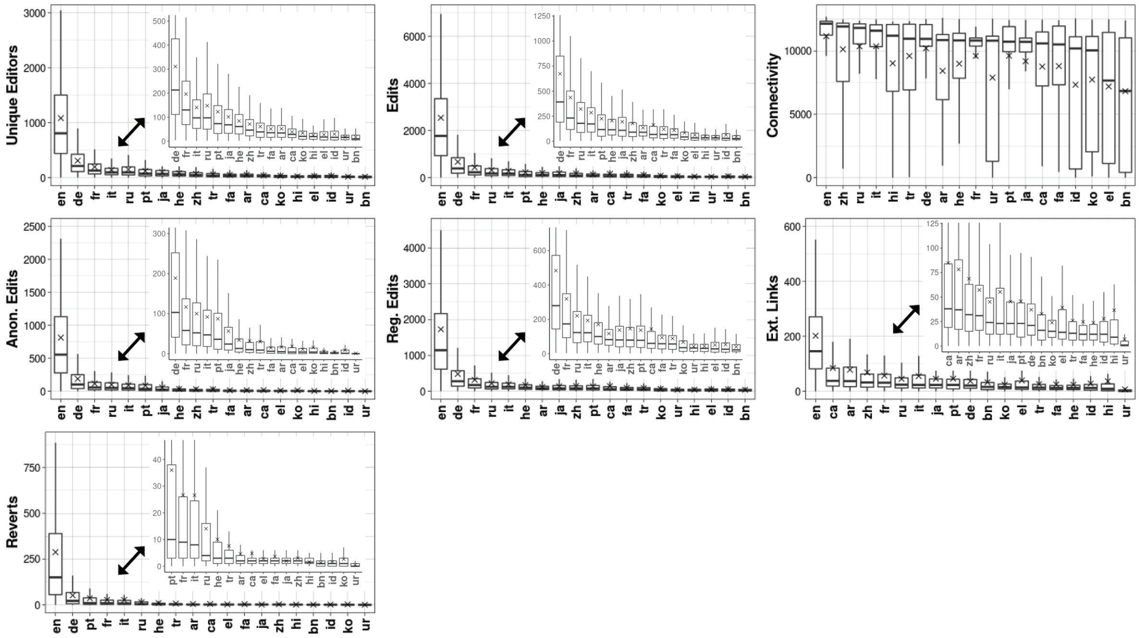


Figure 3: Distributions of authority features

Wikipedia. Registered or anonymous users can make edits, and it is obtained from the article's history.

The Table 3, represents median values for the metric, the interquartile range (IQR), and the idioms with significantly lower quality for each idiom. It is also represented the chi-squared and *p*-value for the metric. In this table, we observe that English stands out from the other idioms, achieving the top score for the median value. English is also the idiom with higher dispersion in values, achieving the higher IQR. German and Russian comes in second and third places. Bengali occupies the last place. Kruskal-Wallis test reveals significant differences among idioms. Russian is, however, not significantly higher than the fourth idiom - Italian. Although Bengali occupies the last place, three idioms - Urdu, Indonesian and Greek - are not significantly higher than Bengali.

Figure 3 represents boxplots for the features distributions. Outliers are not displayed for a more helpful plot visualization, and a zoom layer has been added on the less discernible areas. From this figure, we can conclude that English is the leader in all of the features. The number of reverts varies considerably between idioms, with a very considerable dominance of English. It should be noted that this difference may be due to the fact that there are significant differences in the number of article reversions, but also because that there are reversions that the authors did not identify as such, something that happens mainly in the less developed versions of Wikipedia. In these versions, there is less care with the structure of the articles in general and the comments in particular. English has the largest IQR for all features but connectivity.

5.2 Completeness

Completeness is defined as "the granularity or precision of an information object's model or content values according to some general-purpose IS-A ontology such as WordNet" [29], and it is computed as ***Completeness = 0,4*** $*$ ***Num. Internal Broken Links + 0,4*** $*$ ***Num. Internal Links + 0,2*** $*$ ***Article Length***. Broken links correspond to those linking to pages that are no longer works. Internal links are those referring to internal pages of Wikipedia. The length corresponds to the number of characters of the article's text.

From Table 4, we can conclude that English emerges as the clear leader among the idioms for completeness metric, followed by German and French. Idioms such as Urdu, Korean, and Chinese stand out negatively, scoring more than ten times less than English. English is, once again, the idiom with the more significant variability for the metric. Kruskal-Wallis test reveals significant differences between idioms for this metric. French is, however, not significantly different from the following idiom - Russian. Korean, in penultimate place, is not significantly different from the last classified - Urdu.

As for the features, from Figure 4, we can observe that English scores the highest quality in all of them, but internal broken links, where it gets the least score and Persian reaches the top score. When we cross the number of internal broken links with the number of internal links, we find that, generally, the idioms with the highest number of internal links have the highest number of broken links. English is, however, an exception because despite being the idiom that has the highest number of internal links, it is the one that has the fewest number of broken links. For article length, English gets more than 150% more median value than the second idiom - German and almost 2,200% than Urdu, the last idiom.

Table 4: Idioms quality assessment for completeness metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	5132.4	5420.3	de	fr	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur	18
German	de	3172.8	4705.0	fr*	ru	it	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur		17
French	fr	2619.3	4362.4	it*	ar	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur				15
Russian	ru	2154.0	3262.3	ar*	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur					14
Italian	it	2042.0	3176.2	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Arabic	ar	1810.8	2799.0	pt	el	he	ca	ja	fa	hi	bn	id	tr	zh	ko	ur						13
Portuguese	pt	1471.4	2244.2	he*	ca	ja	fa	hi	bn	id	tr	zh	ko	ur								11
Greek	el	1203.6	2319.7	ja	fa	hi*	bn	id	tr	zh	ko	ur										9
Hebrew	he	1080.2	1511.2	ja*	fa*	bn	id	tr	zh	ko	ur											8
Catalan	ca	936.0	1853.8	bn	id	tr	zh	ko	ur													6
Japanese	ja	903.2	1346.4	id*	tr*	zh	ko	ur														5
Persian	fa	868.4	1179.3	id*	tr*	zh	ko	ur														5
Hindi	hi	830.9	2350.6	id*	tr*	zh	ko	ur														5
Bengali	bn	676.7	995.9	ko	ur																	2
Indonesian	id	629.8	1312.4	ko	ur																	2
Turkish	tr	625.6	1083.5	ko	ur																	2
Chinese	zh	583.0	846.6	ko	ur																	2
Korean	ko	301.2	534.7																			0
Urdu	ur	271.3	422.6																			0
χ^2		4410.6																				
***p*-value**		<2.2e-16																				

* significance level $0.001<p\leq0.05$, significance level $p\leq0.001$ for the rest of the values

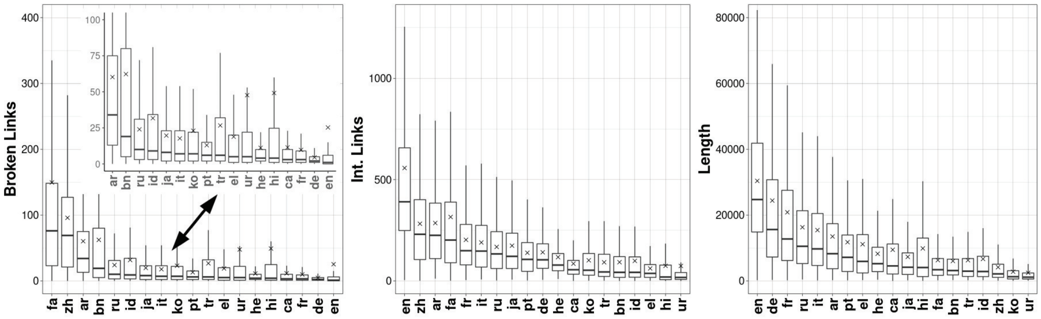


Figure 4: Distributions of completeness features

5.3 Complexity

The definition of complexity is linked to "the degree of cognitive complexity of an information object relative to a particular activity" [29], and it is computed as: ***Complexity = 0,5 ∗ "Flesch Reading Ease" - 0,5 ∗ "Kincaid grade level"***. Both Flesch Reading Ease [11] and Kincaid grade level [16] are tests that assess readability through the number of phrases, words, and syllables of the text. In the Flesch Reading Ease, higher scores mean that the text is easier to read, and lower scores mean that the text is complicated to understand; it is based on a ranking scale of 0-100. The result from Kincaid grade level reflects an American school grade level required to understand the text. They are inversely correlated, as a lower score on the reading ease test corresponds to a higher grade level. These instruments were developed for the English idiom, and so, there are issues when adapting to other idioms. As adaptations of the scales for other idioms are scarce, we decided to consider only the Flesch Reading Ease to calculate this measure. This formula for informativeness was present in an earlier version of Stvilia *et al.* [29]. For the Portuguese, there is an adaptation, consisting of adding 42 to the Flesch Reading Ease [23]; for the Turkish, the Atesman Formula [8] was used; for the English, French, German, and Italian idioms, Flesch Reading Ease was obtained using the Textstat [12] library for Python coding language. For the rest of the idioms, no satisfactory implementation was found.

In the Table 5 are only shown results for the idioms where the Flesch Reading Ease was computed. We can see that Italian and English stand out negatively, and Portuguese positively. Italian is significantly lower than all the idioms, and it gets only 17% of the Portuguese score. Regardless of being the top scorer, the Portuguese does not have the higher IQR. The outliers correspond mainly to articles in the different idioms, consisting of lists, resulting in a wrong result when applying the Flesch Reading Ease index, for example, the list of dead people by COVID-19. When analyzing this metric, it is always necessary to consider that Flesch

Reading Ease was initially developed for English and that the application in other idioms is an adaptation of this metric. Kruskal-Wallis test reveals significant differences among idioms for complexity. Portuguese, the top classified, is significantly higher than all idioms.

Table 5: Idioms quality assessment for complexity metric

		Median	IQR	Significantly lower idioms	# idioms
Portuguese	pt	76	18	tr de* fr en it	5
Turkish	tr	62	34	de en it	3
German	de	56	10	fr en it	3
French	fr	45	18	en* it	2
English	en	34	14	it	1
Italian	it	13	26		0
χ^2		11373			
***p*-value**		<2.2e-16			

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

The only feature considered in the computation of complexity was the Flesch Reading Ease. The respective distribution is shown in Figure 5.

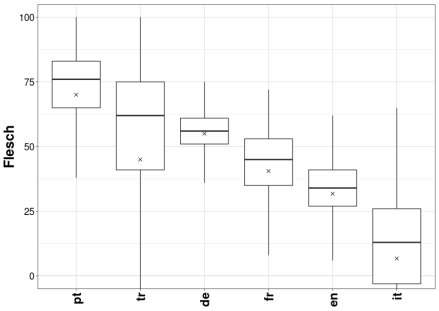


Figure 5: Distribution of complexity feature

5.4 Informativeness

Informativeness is defined as "the amount of information that an information object contains" [29], and it is computed as: ***Informativeness = 0,6 ∗ InfoNoise - 0,6 ∗ Diversity + 0,3 ∗ Num. Images***. InfoNoise [45] refers to the ratio between the size of the information after stemming and stopping and the article size before processing. Diversity corresponds to the ratio between the number of unique editors and the number of edits of an article. The number of images is obtained in the article.

In the Table 6, we can observe that English once again stands out positively, getting more than triple the score of the second idiom - Chinese, in second place. English has the more considerable variability for its median values. Urdu is once more in the last position. Kruskal-Wallis test reveals significant differences among idioms for informativeness. English is significantly higher than all the other idioms, and Urdu, the in the last position, is significantly lower than all the other idioms. Indonesian, Persian, Turkish, and Hindi are all solely significantly higher than Urdu.

Analyzing the features, according to Figure 6, English is just on top for images, getting 20 times the Hindi and Urdu median scores. For diversity, where a lower score means higher quality, Portuguese occupies first place. Japanese get the top score for infoNoise, where Chinese stands out negatively, getting the last place. We can observe similar dispersion for most of the idioms values in all the features.

5.5 Consistency

Consistency is defined as "the extent to which similar attributes or elements of an information object are consistently represented with the same structure, format and precision" [29], and it is computed as: ***Consistency = 0,6 ∗ Administrators Edit Share + 0,5 ∗ Age***. The administrators edit share corresponds to editions made by administrators, and it is obtained in the history. The item's age corresponds to the time difference, in days, between the collection date and the article's creation date.

From Table 7, we can observe that English, German and French again come out positively. English is, once again, the idiom with a higher IQR. Bengali gets the last place for this metric. The distribution is more homogeneous in this metric than in those previously analyzed - the top score is less than four times higher than the last idiom. Outliers refer to the various articles, in different idioms, recent and without any editing by administrators - administrator share is zero. Kruskal-Wallis test reveals significant differences among idioms for consistency. There are, however, no significant differences between English and German, the first and second idioms. There are also no significant differences for Bengali, the last, and Greek, the penultimate classified.

Analyzing the features distribution, from Figure 7, we can observe that English is expectedly at the top in the age feature, as it is the eldest version of Wikipedia. German, the second oldest version of Wikipedia, occupies the second place, but the third most old version - Catalan occupies the eleventh place. French occupies third place, and Bengali occupies once again the last place. In the administrator share, the dominance of the English is very significant, followed by Bengali and Arabic. The share of editions made by administrators is generally low, and for 11 idioms, the median value equals zero, although the mean values are higher than zero. There is a high dispersion in the age feature values for the majority of the idioms.

5.6 Volatility

Volatility is defined as "the amount of time the information remains valid" [29]. It corresponds to the length of hours the content remained valid until a later edition reverted it.

Analyzing Table 8 and Figure 8, we can observe that metric distribution does not follow the same pattern as the previous metrics. The top scores belong to Bengali, Catalan, Indonesian, Korean, and Urdu, whose median values equal zero. English comes in fifth place, and Japanese occupies the last place, as smaller scores translate into higher quality, as a lower median revert time means faster recovery from erroneous editions. However, the volatility score has a singularity - when there are no reversions in articles, the score for volatility equals zero, the same score as an article where the median time for its reversions is zero. This situation is also verified in the works of Domingues and Teixeira Lopes [7] - for the Portuguese idiom and Stvilia *et al.* [30] - for the random dataset. Cross-referencing this data with the number of reverts, we can

Table 6: Idioms quality assessment for informativeness metric

		Median	IQR	Significantly lower idioms																		# idioms
English	en	12.38	19.48	zh	ar	fr	it	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur	18
Chinese	zh	3.72	7.44	ar	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur				15
Arabic	ar	3.53	2.40	fr*	it	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur			16
French	fr	3.01	3.20	it*	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur				15
Italian	it	2.57	1.83	ja	he	ca	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur					14
Japanese	ja	2.08	5.14	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur								11
Hebrew	he	2.01	1.03	ru	de	pt	bn	ko	el	id	fa	tr	hi	ur								11
Catalan	ca	1.83	1.78	de*	bn*	ko*	el	id	fa	tr	hi	ur										9
Russian	ru	1.68	1.34	ko	el	id	fa	tr	hi	ur												7
German	de	1.67	1.88	ko	el	id	fa	tr	hi	ur												7
Portuguese	pt	1.65	1.84	ko	el	id	fa	tr	hi	ur												7
Bengali	bn	1.45	2.87	ko*	id	fa	tr	hi	ur													6
Korean	ko	1.38	0.90	fa*	tr	hi*	ur															4
Greek	el	1.25	1.37	id*	fa	tr	hi	ur														5
Indonesian	id	1.12	1.14	ur																		1
Persian	fa	1.10	1.33	ur																		1
Turkish	tr	1.07	0.81	ur*																		1
Hindi	hi	0.82	1.61	ur*																		1
Urdu	ur	0.72	1.03																			0
χ^2		4446.6																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

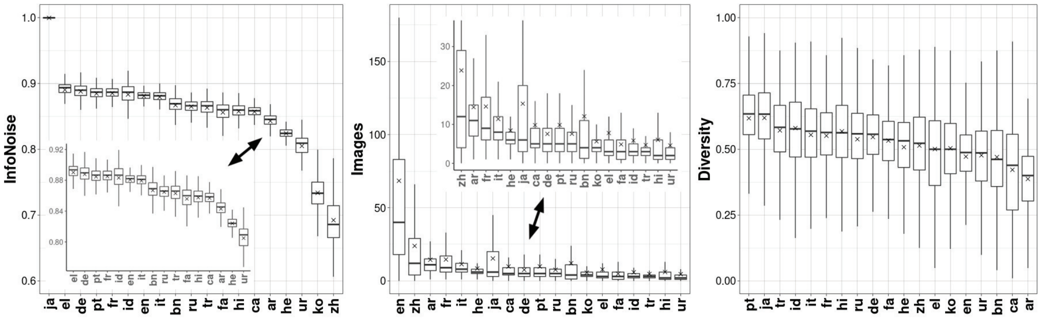


Figure 6: Distributions of informativeness features

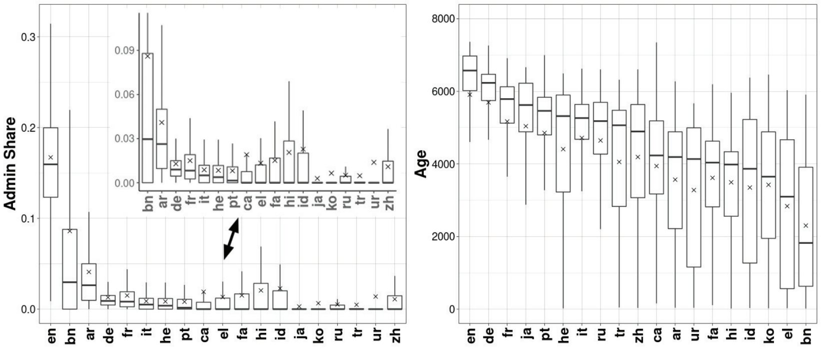


Figure 7: Distributions of consistency features

observe that, for English, we have 0.3% of articles without any revert, while this value rises to 36%, 23%, 36%, 28%, and 74%, for Bengali, Catalan, Indonesian, Korean and Urdu, respectively. Also, we can see that Urdu scores zero for the IQR, as the few values it gets are classified as outliers. According to this analysis, we can consider that these five idioms do not achieve, in fact, more quality, for volatility, than English or French, and so, we consider English as the top score idiom. The Kruskal-Wallis test reveals significant differences among idioms for volatility and median revert time, which is the only feature for this metric. There are, however, no significant differences for the two last classified idioms - Arabic and Japanese.

Table 7: Idioms quality assessment for consistency metric

		Median	IQR	Significantly lower idioms																	# idioms
English	en	3286.4	476.6	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
German	de	3116.5	362.0	fr	ja	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn	17
French	fr	2893.5	501.2	pt	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Japanese	ja	2811.5	669.5	pt*	he	it	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn			15
Portuguese	pt	2731.0	518.5	he*	ru	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn					13
Hebrew	he	2659.0	1336.5	tr	zh*	ca	ar	ur	fa	hi	id	ko	el	bn							11
Italian	it	2632.0	480.3	tr	zh	ca	ar	ur	fa	hi	id	ko	el	bn							11
Russian	ru	2590.0	712.7	tr	zh*	ca	ar	ur	fa	hi	id	ko	el	bn							11
Turkish	tr	2532.5	1324.9	ar	ur	fa	hi	id	ko	el	bn										8
Chinese	zh	2447.0	1284.0	ca	ar	ur	fa	hi	id	ko	el	bn									9
Catalan	ca	2117.0	1007.0	ar*	ur*	fa	hi	ko*	el	bn											7
Arabic	ar	2095.0	1333.5	el	bn																2
Urdu	ur	2067.8	1917.5	bn																	1
Persian	fa	2018.8	904.9	el	bn																2
Hindi	hi	1991.8	889.0	bn																	1
Indonesian	id	1933.0	1977.9	el	bn																2
Korean	ko	1824.3	1467.6	el	bn																2
Greek	el	1549.0	2052.0																		0
Bengali	bn	911.3	1639.8																		0
$\chi2$		4534.3																			
***p*-value**		<2.2e-16																			

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

Table 8: Idioms quality assessment for volatility metric

		Volatility	IQR	Significantly lower idioms																# idioms
Bengali	bn	0	63.0	ur	en*	pt*	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Catalan	ca	0	26.0	id*	ko	ur	en*	pt	de*	it	he	hi	ru	fa	tr	el	zh	ar	ja	16
Indonesian	id	0	50.3	ur	en	fr	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
Korean	ko	0	32.0	ur	en	fr	it*	he	hi	ru	fa	tr	el	zh	ar	ja				13
Urdu	ur	0	0.0	fr	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja				13
English	en	2	3.0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
French	fr	4	6.0	pt	de	it	he	hi	ru	fa	tr	el	zh	ar	ja					12
Portuguese	pt	5	18.0	he	hi	ru	fa	tr	el	zh	ar	ja								9
German	de	5.5	12.0	it	he	hi	ru	fa	tr	el	zh	ar	ja							10
Italian	it	6.5	30.0	ru	fa	tr	el	zh	ar	ja										7
Hebrew	he	13	31.0	ru	fa	tr	el	zh	ar	ja										7
Hindi	hi	19	31.8	ru	fa	tr	el	zh	ar	ja										7
Russian	ru	26	77.5	ar*	ja															2
Persian	fa	32.5	87.5	ar	ja															2
Turkish	tr	33	99.0	ar*	ja															2
Greek	el	36	63.0	ar*	ja															2
Chinese	zh	40	58.0	ja																1
Arabic	ar	56	83.0																	0
Japanese	ja	58	90.5																	0
$\chi2$		2775.9																		
***p*-value**		<2.2e-16																		

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

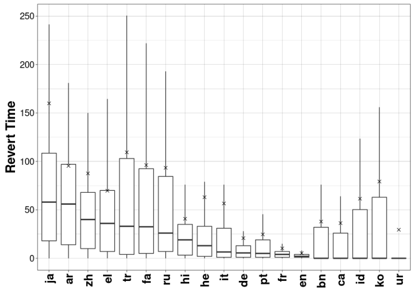


Figure 8: Distribution of volatility feature

5.7 Currency

Currency is "the age of an information object" [29]. It corresponds to the time between the collection date and the date of the article's last update, in days.

In currency, a lower score means higher quality, as lower currency means more up-to-date articles. Analyzing Figure 9 and Table 9, English stands out positively, followed by German and Japanese. English also gets the lowest IQR. Hindi and Urdu stand out negatively. Urdu score is notably lower than the top score (1,785%). The outliers, removed, correspond to various articles in different idioms, where no edition has been made for a long time. Kruskal-Wallis test reveals significant differences among idioms for currency. While English is significantly higher than the rest of the idioms, German is not significantly higher than Japanese. Urdu is significantly lower

Table 9: Idioms quality assessment for currency metric

		Median	**IQR**	**Significantly lower idioms**																		**# idioms**
English	en	20	17	de	ja	ca	ru	fr	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur	18
German	de	42	72	ca	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Japanese	ja	42	39	ca*	fa	tr	it	ar	zh	he	bn	ko	el	id	pt	hi	ur					14
Catalan	ca	50	99	ar	zh	bn	ko	el	id	pt	hi	ur										9
Russian	ru	50	90	fa	tr	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
French	fr	54	82	fa	tr*	it*	ar	zh	he	bn	ko	el	id	pt	hi	ur						13
Persian	fa	57	114	bn*	ko	el	id	pt	hi	ur												7
Turkish	tr	63	66	bn*	ko	el	id	pt	hi	ur												7
Italian	it	65	101	bn	ko	el	id	pt	hi	ur												7
Arabic	ar	71	144	ko	el	id	pt	hi	ur													6
Chinese	zh	75	124	he	ko	el	id	pt	hi	ur												7
Hebrew	he	79	105	bn*	ko	el	id	pt	hi	ur												7
Bengali	bn	87	205	el	id	pt	hi	ur														5
Korean	ko	126	169	el*	id	pt	hi	ur														5
Greek	el	130	256	hi*	ur																	2
Indonesian	id	140	349	hi*	ur																	2
Portuguese	pt	157	283	hi*	ur																	2
Hindi	hi	210	218																			0
Urdu	ur	357	370																			0
χ^2		2802.9																				
***p*-value**		<2.2e-16																				

* significance level $0.001 < p \leq 0.05$, significance level $p \leq 0.001$ for the rest of the values

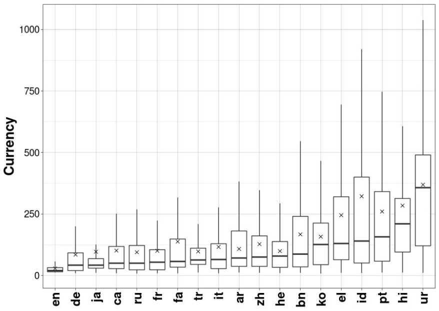


Figure 9: Distribution of currency feature

than the rest of the idioms. The only feature of currency is the currency itself.

6 DISCUSSION

When we analyze the scores of the idioms in the different metrics and their metrics, we see that some idioms generally occupy the top places and others more often occupy the bottom places. In order to be able to define a ranking of idioms, we computed the mean of the number of significantly lower languages for all metrics. These values are present in Table 10, where the idioms are sorted in descending order of the mean since a higher mean means that the language scored significantly higher than the rest. It is also represented the percentile rank for each idiom. Top scores are highlighted in bold. In currency, it is necessary to take into account the constraints associated with Bengali, Catalan, Indonesian, Korean, and Urdu, and described in Subsection 5.6.

As expected, English is at the top, with a mean of significantly lower idioms of 14.6 and a mean percentile of 85%. German and French scored a mean of 12.1, and mean percentiles of 79% and 76%, respectively. The last place belongs to Urdu, with a mean of significantly lower idioms of 2.5 and a mean percentile of 22%. Greek scored not very far from Urdu, with means of 3.0 and 25%, for significantly lower idioms and percentile, respectively. Given the already discussed idiosyncrasies of volatility and considering English as the top idiom, the only metric that does not rank first is complexity, but this metric may be subject to constraints previously described. Regarding the idioms selected for their historical tradition, we can observe that Greek, Persian, Turkish, and Korean are on the bottom half of the table. On the other hand, Italian was the idiom that got the best mean of significantly lower idioms - 9.4, with a mean percentile of 58%.

These results point in the direction of other works, such as Teixeira Lopes and Ribeiro [21], suggesting that English should be provided to users with higher levels of English proficiency, opening doors for higher-quality content.

To understand if quality has any connection with quantity, we computed the correlation between the quality across the different metrics and the number of speakers and the total number of articles in each Wikipedia version for the selected idioms. Results are shown in Table 11, which presents the Spearman correlation value and *p*-values for the number of speakers and Wikipedia articles, for all idioms, and the metrics quality computed values.

Analyzing the results, we can conclude that there is a significant correlation between quality and the number of total articles in each Wikipedia version, mainly for completeness (0.94), authority (0.9), and informativeness (0.9), with significant *p*-values(≤0.001). Informativeness is the metric with more correlation with the number of speakers, with a significant *p*-value (≤0.001), followed by authority (*p*-value≤0.05). There is also a strong correlation (0.63) for the number of speakers and the number of articles in each Wikipedia version, with a significant computed *p*-value(≤0.05).

To analyze the effects of the different number of articles in languages, we have conducted a similar analysis, including only the 164 articles having a version in every studied idiom. From this analysis, we could conclude that the significant results are very

Table 10: Idioms ranking summary

	Authority		**Completeness**		**Complexity**		**Informativeness**		**Consistency**		**Volatility**		**Currency**		**Mean**	
	# SLI	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**# SLI**	**%**	**SLI**	**%**
English	**18**	**100%**	**18**	**100%**	1	33%	**18**	**100%**	**17**	**95%**	12	68%	**18**	**100%**	**14.6**	**85%**
German	17	95%	17	95%	3	67%	7	47%	**17**	**95%**	10	63%	14	89%	12.1	79%
French	13	79%	15	89%	2	50%	15	84%	15	84%	12	68%	13	79%	12.1	76%
Russian	14	84%	14	84%			7	47%	11	63%	2	21%	13	79%	10.2	63%
Italian*	14	84%	13	74%	0	17%	14	79%	11	63%	7	42%	7	47%	9.4	58%
Catalan*	5	37%	6	53%			9	63%	7	47%	**16**	**100%**	9	74%	8.7	62%
Hebrew*	6	53%	8	58%			11	68%	11	63%	7	42%	7	47%	8.3	55%
Japanese	5	37%	5	37%			11	68%	15	84%	0	5%	14	89%	8.3	54%
Portuguese	9	68%	11	68%	**5**	**100%**	7	47%	13	79%	9	58%	2	16%	8.0	62%
Chinese	12	74%	2	16%			15	84%	9	58%	1	16%	7	47%	7.7	49%
Arabic	5	37%	13	74%			16	95%	2	26%	0	5%	6	42%	7.0	46%
Bengali	0	5%	2	16%			6	42%	0	5%	13	79%	5	32%	4.3	30%
Korean*	1	21%	0	5%			4	32%	2	26%	13	79%	5	32%	4.2	32%
Turkish*	6	53%	2	16%	3	67%	1	11%	8	53%	2	21%	7	47%	4.1	38%
Persian*	4	32%	5	37%			1	11%	2	26%	2	21%	7	47%	3.5	29%
Hindi	6	53%	5	37%			1	11%	1	16%	7	42%	0	5%	3.3	27%
Indonesian	0	5%	2	16%			1	11%	2	26%	13	79%	2	16%	3.3	25%
Greek*	0	5%	9	63%			5	37%	0	5%	2	21%	2	16%	3.0	25%
Urdu	1	21%	0	5%			0	5%	1	16%	13	79%	0	5%	2.5	22%

SLI: Significantly lower idioms, * idioms selected for their historical tradition

Table 11: Correlation between metrics and number of speakers and articles

	Speakers		**Wikipedia**	
	correlation	*p*-value	correlation	*p*-value
Authority	0.67	0.0015*	0.90	0.0000**
Completeness	0.46	0.0459	0.94	0.0000**
Complexity	0.06	0.8091	0.37	0.1162
Informativeness	0.77	0.0001**	0.90	0.0000**
Consistency	0.29	0.2247	0.69	0.0011*
Volatility	0.01	0.9597	-0.15	0.5361
Currency	-0.08	0.7544	-0.46	0.0451

* significance level $p \leq 7e\text{-}3$, ** significance level $p \leq 1e\text{-}4$. (Bonferroni corrected from $p=0.05$ and $p=0.001$, 7 tests)

similar to those described above. The top and bottom-ranked languages remain the same, and English is still leading in the same metrics.

7 CONCLUSION

We performed a comparison of health-related articles on Wikipedia across 19 different idioms: English, Arabic, French, Portuguese, German, Persian, Italian, Chinese, Russian, Japanese, Hebrew, Korean, Catalan, Indonesian, Turkish, Greek, Hindi, Bengali, and Urdu. To assess the information quality of the articles, we used a set of seven predefined metrics: authority, completeness, complexity, informativeness, consistency, currency, and volatility.

We faced some challenges due to the heterogeneity of the idioms analyzed and some variation between the different versions of Wikipedia, such as its structure. Given this heterogeneity, we could not use some metrics in some idioms. It is the case of the readability tests for complexity metric.

After analyzing the results, we concluded that there is a significant difference among idioms for quality. English is the idiom that shows the most difference to all other idioms, with the best values for quality metrics, followed by German, French, and Russian. Urdu, Greek, Indonesian, and Hindi are the idioms with worse values of quality in general. We also concluded a correlation between the number of speakers and the number of articles in each Wikipedia version. This correlation is more significant for the number of Wikipedia's articles and for some metrics, such as completeness and authority. With this characterization of the differences between idioms, we hope to raise awareness of this heterogeneity and make the first step towards more equal versions of Wikipedia. To overcome this heterogeneity between the different idiom versions of Wikipedia, the Wikimedia Foundation has an ongoing project - Abstract Wikipedia [36]. This project aims to create a language-independent version of Wikipedia by modeling data from Wikidata. This will allow people to create language-independent content that will be later translated through code. This project also contains Wikifunctions, which allows anyone to create and maintain code and includes code that converts the language-independent article from Abstract Wikipedia to Wikipedia's native language.

REFERENCES

[1] Rita Baeten, Slavina Spasova, Bart Vanhercke, and Stéphanie Coster. 2018. *Inequalities in access to healthcare.* Number November.

[2] Anamika Chhabra, Shubham Srivastava, S Iyengar, and Poonam Saini. 2021. Structural Analysis of Wikigraph to Investigate Quality Grades of Wikipedia Articles. https://doi.org/10.1145/3442442.3452345

[3] Riccardo Conti, Emanuel Marzini, Angelo Spognardi, Ilaria Matteucci, Paolo Mori, and Marinella Petrocchi. 2014. Maturity assessment of Wikipedia medical articles. *Proceedings - IEEE Symposium on Computer-Based Medical Systems* (2014), 281–286. https://doi.org/10.1109/CBMS.2014.69

[4] Luís Couto and Carla Lopes. 2021. Assessing the quality of health-related Wikipedia articles with generic and specific metrics. 640–647. https://doi.org/10.1145/3442442.3452355

[5] Baptiste de La Robertie, Yoann Pitarch, and Olivier Teste. 2015. Measuring article quality in Wikipedia using the collaboration network. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015* (2015), 464–471. https://doi.org/10.1145/2808797.2808895

[6] Lara Devgan, Neil Powe, Brittony Blakey, and Martin Makary. 2007. Wiki-Surgery? Internal validity of Wikipedia as a medical and surgical reference. *Journal of the American College of Surgeons* 205 (09 2007), S76–S77. https://doi.org/10.1016/j.jamcollsurg.2007.06.190

[7] Gil Domingues and Carla Teixeira Lopes. 2019. Characterizing and comparing Portuguese and English Wikipedia medicine-related articles. *The Web Conference*

2019 - Companion of the World Wide Web Conference, WWW 2019 (2019), 1203–1207. https://doi.org/10.1145/3308560.3316758

[8] Ateşman E. 1997. Measuring readability in Turkish. *Tömer Language Journal* 58 (1997), 171–174.

[9] Ethnologue. 2021. How many languages are there in the world? Retrieved jun, 2021 from https://www.ethnologue.com/guides/how-many-languages

[10] Elena Filatova. 2009. Directions for Exploiting Asymmetries in Multilingual Wikipedia. (01 2009). https://doi.org/10.3115/1572433.1572438

[11] R FLESCH. 1948. A new readability yardstick. *The Journal of applied psychology* 32, 3 (June 1948), 221—233. https://doi.org/10.1037/h0057532

[12] Python Software Foundation. 2021. textstat 0.7.0. Retrieved jun, 2021 from https://pypi.org/project/textstat/

[13] Google. 2021. Year in Search 2020. Retrieved jun, 2021 from https://trends.google.com/trends/yis/2020/GLOBAL/

[14] Scott Hale. 2013. Multilinguals and Wikipedia Editing. *WebSci 2014 - Proceedings of the 2014 ACM Web Science Conference* (12 2013). https://doi.org/10.1145/2615569.2615684

[15] Imran Khan, Shahid Hussain, Hina Gul, Muhammad Shahid, and Muhammad Jamal. 2019. *An Empirical Study to Predict the Quality of Wikipedia Articles.* 485–492. https://doi.org/10.1007/978-3-030-16187-3_47

[16] J. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.

[17] Jona Kräenbring, Tika Penza, Joanna Gutmann, Susanne Muehlich, Oliver Zolk, Leszek Wojnowski, Renke Maas, Stefan Engelhardt, and Antonio Sarikas. 2014. Accuracy and Completeness of Drug Information in Wikipedia: A Comparison with Standard Textbooks of Pharmacology. *PloS one* 9 (09 2014), e106930. https://doi.org/10.1371/journal.pone.0106930

[18] Michaël R. Laurent and Tim J. Vickers. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16, 4 (2009), 471–479. https://doi.org/10.1197/jamia.M3059

[19] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2020. Modeling Popularity and Reliability of Sources in Multilingual Wikipedia. *Information* 11 (05 2020), 263. https://doi.org/10.3390/info11050263

[20] Xinyi Li, Jintao Tang, Ting Wang, Zhunchen Luo, and Maarten de Rijke. 2015. Automatically assessing wikipedia article quality by exploiting article–editor networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9022 (2015), 574–580. https://doi.org/10.1007/978-3-319-16354-3_64

[21] Carla Teixeira Lopes and Cristina Ribeiro. 2013. Measuring the value of health query translation: An analysis by user language proficiency. *JASIS* 64, 5 (2013), 951–963. https://doi.org/10.1002/asi.22812

[22] Edison Marrese-Taylor, Pablo Loyola, and Yutaka Matsuo. 2019. An Edit-centric Approach for Wikipedia Article Quality Assessment. (2019), 381–386. https://doi.org/10.18653/v1/d19-5550 arXiv:1909.08880

[23] Teresa Martins, Claudete Ghiraldelo, M. Gragas, V. Nunes, and O.N. Oliveira Jr. 1996. Readability Formulas Applied to Textboooks in Brazilian Por- tuguese. (06 1996).

[24] Sorin Matei and Brian Britt. 2017. *Structural Differentiation in Social Media: Adhocracy, Entropy, and the "1% Effect".* https://doi.org/10.1007/978-3-319-64425-7

[25] Omeed Modiri, Daipayan Guha, Naif M. Alotaibi, George M. Ibrahim, Nir Lipsman, and Aria Fallah. 2018. Readability and quality of wikipedia pages on neurosurgical topics. *Clinical Neurology and Neurosurgery* 166, January (2018), 66–70. https://doi.org/10.1016/j.clineuro.2018.01.021

[26] Daniel Pimienta, D. Prado, and Á Blanco. 2009. Twelve years of measuring linguistic diversity in the Internet: balance and perspectives. *Paris: UNESCO. Retrieved March* 7, September (2009), 2010.

[27] Natalia Pletneva, Sarah Cruchet, Maria Ana Simonet, Maki Kajiwara, and Célia Boyer. 2011. Results of the 10th HON survey on health and medical internet use. *Studies in Health Technology and Informatics* 169, 2008 (2011), 73–77. https://doi.org/10.3233/978-1-60750-806-9-73

[28] Malolan Rajagopalan, Vineet Khanna, Yaacov Leiter, Meghan Stott, Timothy Showalter, Adam Dicker, and Yaacov Lawrence. 2011. Patient-Oriented Cancer Information on the Internet: A Comparison of Wikipedia and a Professionally Maintained Database. *Journal of oncology practice / American Society of Clinical Oncology* 7 (09 2011), 319–23. https://doi.org/10.1200/JOP.2010.000209

[29] Besiki Stvilia, Michael Twidale, Linda Smith, and Les Gasser. 2005. Assessing Information Quality of a Community-Based Encyclopedia. *Proceedings of the 2005 International Conference on Information Quality, ICIQ 2005* (01 2005).

[30] B. Stvilia, M. B. Twidale, L. Gasser, and L. C. Smith. 2005. Information quality in a community-based encyclopedia. *Knowledge Management: Nurturing Culture, Innovation, and Technology-Proceedings of the 2005 International Conference on Knowledge Management* (2005), 101–113.

[31] Athikhun Suwannakhan, Daniel Casanova-Martínez, Laphatrada Yurasakpong, Punchalee Montriwat, Krai Meemon, and Taweetham Limpanuparb. 2019. The Quality and Readability of English Wikipedia Anatomy Articles. *Anatomical Sciences Education* 13 (2019), 1–13. https://doi.org/10.1002/ase.1910

[32] Sharon Tan and Nadee Goonawardene. 2017. Internet Health Information Seeking and the Patient-Physician Relationship: A Systematic Review. *Journal of Medical Internet Research* 19 (01 2017), e9. https://doi.org/10.2196/jmir.5729

[33] Garry R. Thomas, Lawson Eng, Jacob F. de Wolff, and Samir C. Grover. 2013. An Evaluation of Wikipedia as a Resource for Patient Education in Nephrology. *Seminars in Dialysis* 26, 2 (2013), 159–163. https://doi.org/10.1111/sdi.12059

[34] Ziko VanDijk. 2009. Wikipedia and lesser-resourced languages. *Language Problems and Language Planning* 33, 3 (2009), 234–250. https://doi.org/10.1075/lplp.33.3.03van

[35] Peter Volsky, Cristina Baldassari, Sirisha Mushti, and Craig Derkay. 2012. Quality of Internet information in pediatric otolaryngology: A comparison of three most referenced websites. *International journal of pediatric otorhinolaryngology* 76 (07 2012), 1312–6. https://doi.org/10.1016/j.ijporl.2012.05.026

[36] Wikimedia. 2021. Abstract Wikipedia. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Abstract_Wikipedia

[37] Wikimedia. 2021. Language proposal policy. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Language_proposal_policy

[38] Wikimedia. 2021. Requests for new languages. Retrieved jun, 2021 from https://meta.wikimedia.org/wiki/Requests_for_new_languages

[39] Wikipedia. 2021. List of Wikipedias. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/List_of_Wikipedias

[40] Wikipedia. 2021. Wikipedia:Blocking policy. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Blocking_policy

[41] Wikipedia. 2021. Wikipedia:Content assessment. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Content_assessment

[42] Wikipedia. 2021. Wikipedia:Five pillars. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

[43] Wikipedia. 2021. Wikipedia:WikiProject Medicine/Popular pages. Retrieved jun, 2021 from https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine/Popular_pages

[44] Kewen Wu, Qinghua Zhu, Yuxiang Zhao, and Hua Zheng. 2010. Mining the factors affecting the quality of Wikipedia articles. *Proceedings - 2010 International Conference of Information Science and Management Engineering, ISME 2010* 1, 1 (2010), 343–346. https://doi.org/10.1109/ISME.2010.114

[45] Xiaolan Zhu and Susan Gauch. 2000. Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Athens, Greece) *(SIGIR '00)*. Association for Computing Machinery, New York, NY, USA, 288–295. https://doi.org/10.1145/345508.345602