Data Brief: Twenty years of offline meeting data of the German-language Wikipedia

Nicole Schwitter
University of Warwick

Abstract

This data brief will present the *dewiki meetup dataset* which covers all 4418 meetups organized on the German-language Wikipedia with information on attendees, apologies, date and place of meeting, and minutes recorded. It is a valuable source of data for Wikimedia and (computational) social science research generally.

Keywords: German Wikipedia, offline meetings, face-to-face meetings, dataset, network data

Introduction

In today's internet landscape, the online encyclopaedia Wikipedia takes on a key role and has developed into a phenomenon met with strong interest by the scientific community. Throughout the past 20 years of its existence, Wikipedia and its specific software structure have created a rich and freely accessible data source on various online activities undertaken, offering the opportunity to study large-scale, self-organizing collaboration networks. This has made Wikipedia a popular source of data for scholars of various disciplines. However, Wikipedia also has a notable offline component which has been largely neglected by the scientific community. Wikipedia is characterized by regular local offline meetups, which give editors a time and place to get to know each other personally. These meetings are organized publicly and are well-documented with lists of attendees and minutes. This data brief presents the newly published *dewiki meetup dataset* containing all meetups organized on the German-language Wikipedia. From Wikipedia's launch in 2001 to March 2020 when face-to-face meetings came to a halt due to the outbreak of the Coronavirus pandemic, all meetups were collected and can now be merged with online activity data. The dataset forms a valuable source of data for Wikimedia researchers and social science research generally: It captures the development of the offline network over time of one of the most sustainable, online public goods and the community producing it.

Collection of Meetup Data

Data on offline meetings is generally publicly available on Wikipedia; however, the data lacks a clear and consistent structure as it is user-written. The starting point of the meetup collection was an overview list[1] of meetings between Wikipedians. Additionally, all editathons, open editing events and general events listed on different overview sites were collected. Lastly, all WikiProjects and task forces were checked for meetings. Throughout the scraping of all these pages, a snowballing approach was followed.

Some pages and meetings were excluded from the data and/or the data collection process. First, all meetups that took place only virtually were skipped. Next, portals (introductory landing pages for readers) were not checked for meetups unless they are covering regional entities. This dataset is further restricted to meetings organized on the German-language version of Wikipedia. All meetings not organized on the German-language Wikipedia (but instead, for example, on meta or commons) were excluded from data collection. Additionally, very regular meetings taking place in community spaces were excluded. Community spaces are places of often extraordinarily high activity and meetings are often attended by the same very small, core group of editors which stop recording their attendance. As these community spaces thus exhibit a very different dynamic and as it was often impossible to reliably collect data on the attendees, such meetups were excluded from the dataset.

The data collection aimed to collect information on at least the date, place/venue, and attendees of all offline meetings. In most cases, the data collected also included apologies for absences and minutes about the meetup. When possible, an automatic scraper was written to extract the information, but most meeting data was collected manually.

Collected Data

The *dewiki meetup dataset* includes 4418 meetings. Full documentation of all variables is available with the dataset.

The first meeting recorded took place on October 28th 2003 with five attendees in Munich, the last ones on March 13th 2020 with three attendees in Cologne and with four attendees in Leipzig. 77.0 per cent of those 4418 meetups are classified as mainly social while the other 23.0 per cent are considered work meetings (meetings during which attendees more directly contribute towards Wikipedia/Wikimedia). The

[1] https://de.wikipedia.org/wiki/Kategorie:Wikipedia:Treffen_der_Wikipedianer.

distribution of meetups over time is pictured in figure 1 (for 2020, a total of 67 meetups are in the dataset with 38 being social in nature; 56.7 per cent). The spatial global distribution of meetups is plotted in figure 2. The large majority of meetups, 88.8 per cent (3922), took place in Germany, 5.5 per cent (244) in Austria, 4.3 per cent (188) in Switzerland and 0.023 per cent (1) in Liechtenstein. Even though this captures around 99 per cent of the meetups, the remaining per cent took place in twenty different countries.

Through offline meetups, a network develops: Users attending the same meetings get to know each other and develop a tie. This leads to an affiliation network of users belonging to meetups. Figure 3 shows the network at different points in time; the plot uses the same color scheme before, thus capturing geographical clusters. The network started in October 2003 with the first meetup in Munich. It then consisted of one cluster with a meeting and its five attendees. By the end of the year, one other meetup took place, giving an additional user the chance to join the meetup scene. By the end of 2005, there were 537 nodes in the network, consisting of 413 users who attended 124 different meetings and belonging to one large and one small component. Some geographical clustering is still observable. At the end of data collection in 2020, the meeting-to-user network consists of 8540 nodes (with 4122 users who have attended 4418 meetings) and becomes too crowded for a meaningful, graphical representation. The network features five clusters; most nodes belong to one large component, while the four other components are small, single meetups with eight, five, and two (2x) attendees, respectively.

Using the Data

The offline meetup data can be combined with online activity data available on Wikipedia. To combine *the dewiki meetup dataset* with the rich body of offline data, it needs to be merged on the basis of usernames (which can change). To account for this, all Wikipedia users and the redirection links linking to them were collected and all re-names as logged in the renaming logbook were web-scraped. This led to a list of 1’751’808 different usernames (and variants of their encoding and spelling) belonging to 1’149’511 unique IDs. This information is saved within the *dewiki meetup data*, as file *nametoid*. This list can be merged with the usernames to replace usernames with IDs.

The *dewiki meetup dataset* can be accessed via the OpenScienceFramework: https://doi.org/10.17605/OSF.IO/EHA4R (Schwitter, 2023a) (CC BY-SA 4.0). The dataset opens the door to novel research opportunities. It provides data on a complete social network: All edges between all users are recorded. The potential of the *dewiki meetup dataset* is unleashed when it is merged with online behavior data. It can then be used to bridge the gap between offline and online actions. The strengths of the dataset also come with some privacy concerns that must be kept in mind. For example, geographical information is recoverable from the dataset and can be linked to users. It is important to use the data fairly and reasonably.

In its first applications, I used the data to discuss the meetups' (causal) effects on contributions (for an early use of parts of the meetup data on this, see Stegbauer, 2009: Chapter 15) and to explore how offline social ties matter in the context of explaining participation in governance activities (Schwitter, 2022, 2023b). Future research can now build upon to explore more complex and nuanced questions. For example, the dataset can be used to study the structure and patterns of interactions among Wikipedia editors. The network of face-to-face interactions can be analyzed to identify the most central and influential editors and the formation of cliques or subgroups. As the data covers a time span of twenty years and dozens of regional communities, it can also be studied how these patterns change over time and depend on the local context. It is also of interest to understand how offline meetings work as places of socialization where community norms are potentially learnt. To extend previous work, researchers could analyze whether articles that were edited by Wikipedians who attended offline meetups have higher levels of accuracy, completeness, or readability than articles that were not edited by those editors. More generally, it can be studied if and how a user’s way of contributing towards Wikipedia changes or improves after attending a meetup. Overall, the *dewiki meetup dataset* presents a unique opportunity for researchers to study the dynamics of online collaboration and the role of social networks in the functioning of online communities.

References

Schwitter Nicole. 2022. *The Role of Offline Ties in Online Communities: The Case of Wikipedia*. PhD Thesis.

Schwitter Nicole. 2023a. Bridging the offline and the online: Twenty years of offline meeting data of the German-language Wikipedia. *SocArXiv*. DOI: 10.31235/osf.io/g96tk.

Schwitter Nicole. 2023b. If I know you offline, I will vote for you online? The role of offline ties in an online public election. *SocArXiv*. DOI: 10.31235/osf.io/dnrha.

Stegbauer Christian. 2009. *Wikipedia*. Springer VS Verlag für Sozialwissenschaften.

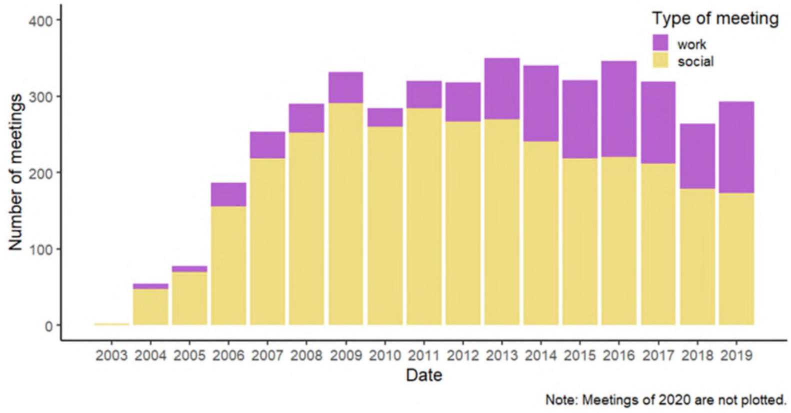

Figure 1: Temporal distribution of meetups

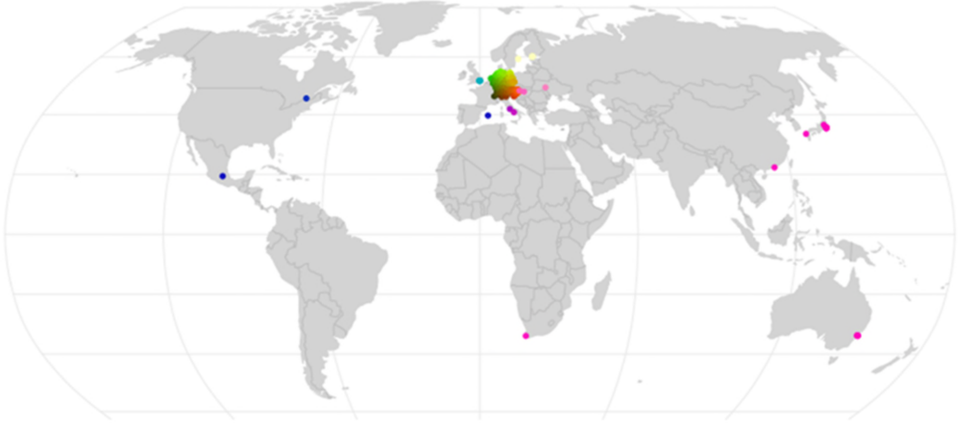

Figure 2: Spatial distributions of meetups. The points are colored according to their longitude and latitude with meetups being close to each other being similar in color (normalized to the DACH region)

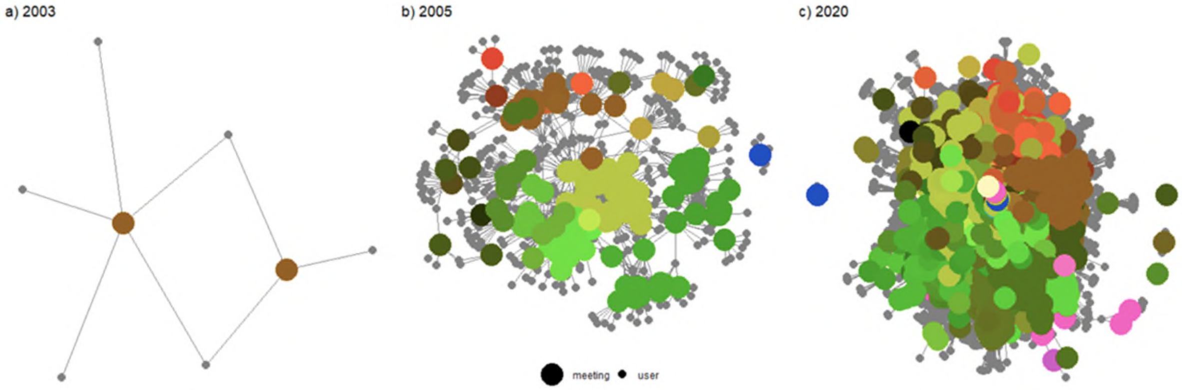

Figure 3: Meeting-to-user network in a) 2003, b) 2005, and c) 2020. The points are colored according to their longitude and latitude with meetups being close to each other being similar in color (normalized to the DACH region)