

D2.2 Social interaction dataset

*Multilingual dataset of social interactions around content
related to the European elections in Wikipedia*

Table of content

Introduction.....	2
Data collection.....	4
Identification of relevant pages.....	4
Articles and categories.....	4
Wikidata and article lists.....	5
Language selection.....	5
Interaction data retrieval.....	6
Dataset description.....	7
Dataset general statistics.....	7
Temporal evolution.....	10
Activity distribution.....	14
Cross-language coverage.....	16
Conclusions.....	19
References.....	20

Introduction

The DEM-Debate project is focused on community-governed platforms, to study their practices based on a distributed and transparent model of creating and sharing information online, and learn about what works and what does not to improve transparency and resilience of the online information ecosystem.

The project in particular focuses on community practices for ensuring information reliability during the electoral campaign for the European elections of 2024, and on Wikipedia as a paradigmatic and especially relevant case of community-governed platforms. Wikipedia is the largest existing free knowledge platform and the only non-commercial VLOP (Very Large Online Platforms)¹ designated under the DSA (Digital Service Act)², visited daily by millions of users around the globe, and edited by thousands, in over 300 different languages.

Wikipedia's content constitutes an important milestone of today's internet, not only as it often appears among the top search results in search engines for queries about almost any kind of topics (Vincent & Hecht, 2021), and is widely shared through other online platforms and media, but more recently also as a repository large employed to train large language models (McDowell, 2024).

Contrary to commercial platforms, and in particular the other platforms designated as VLOPs by the DSA, that have increasingly higher barriers preventing researchers from accessing their data about social interactions and algorithmic decisions, Wikipedia's open model based on principles of collaboration and transparency includes making the whole history of interactions behind the creation of its content openly available to anybody.

The full availability of the edit history for all Wikipedia pages allows us to run a computational analysis of the community interactions in articles and talk pages of Wikipedia encyclopaedic entries related to the European elections of 2024, and to study community dynamics, controversies, conflict resolution and practices with respect to information reliability during the electoral campaign (Laniado et al, 2024).

The first step for performing the computational analysis (task T2.5) is the data collection (Task T2.2 - Collection of digital traces), that is the subject of this deliverable; this includes identifying relevant Wikipedia pages in selected language editions, and retrieving the history of interactions and associated metadata for these pages, strictly adhering to the guidelines and ethical principles detailed in Deliverable D1.4 - Data Management Plan.

¹ https://ec.europa.eu/commission/presscorner/detail/en/ip_23_2413/smo

² <http://data.europa.eu/eli/reg/2022/2065/oj/eng>

This document is structured as follows: in the next section we describe the approach and the process we followed for data collection; then in the Dataset description section we provide some general statistics about the data collected, and a first high level description including aspects such as temporal evolution, activity distribution and cross-language coverage of content; finally in the last section we draw some conclusions.

Data collection

In this section we explain how we designed and executed the data collection process, identifying a list of relevant articles and associated talk pages from all the Wikipedia language editions considered for the project, and retrieving their edit history from the Wikipedia APIs.

Identification of relevant pages

To identify pages related to the European elections of 2024 and the candidates and parties involved from the different European countries and in different languages, we used a combined approach based on article lists and categories from the English Wikipedia, and Wikidata queries (Miquel-Ribé & Laniado, 2019).

Once a list of articles from the English Wikipedia has been identified, it is possible to find the corresponding articles in other relevant languages, when they exist, as links between representations of the same articles in different language editions (interlanguage links³) are provided by the community. The rationale for following this approach is that the most relevant political actors normally have a page in the English Wikipedia which de facto a kind of lingua franca on the internet and in Wikipedia as well as in the European Union; therefore, although some specific articles may exist only in some other languages, we assume that the English Wikipedia includes all the most relevant political actors involved in the European elections, including all the elected MEPs and political parties running for the European elections in the different members countries of the European Union. This way of collecting data allows us to have a uniform and replicable criterion for all countries, guaranteeing that the most relevant political actors are included.

Articles and categories

We started from a core of a few articles strongly related to the European elections:

https://en.wikipedia.org/wiki/2024_European_Parliament_election

https://en.wikipedia.org/wiki/Political_groups_of_the_European_Parliament

https://en.wikipedia.org/wiki/European_political_party

https://en.wikipedia.org/wiki/Politics_of_the_European_Union

https://en.wikipedia.org/wiki/Opinion_polling_and_seat_projections_for_the_2024_European_Parliament_election

³ https://en.wikipedia.org/wiki/Help:Interlanguage_links

We then collated a list of core articles related to the European elections of 2024 from the corresponding category on the English Wikipedia:

https://en.wikipedia.org/wiki/Category:2024_European_Parliament_election

Further we added the articles included in other categories strictly related to the elections:

https://en.wikipedia.org/wiki/Category:Political_groups_of_the_European_Parliament

https://en.wikipedia.org/wiki/Category:European_political_alliances

https://en.wikipedia.org/wiki/Category:Parties_represented_in_the_European_Parliament

Wikidata and article lists

Wikidata⁴ is a collaborative knowledge base associated with Wikipedia and its Wikimedia sister projects, that provides a central store of structured data edited and maintained by both humans and machines. From Wikidata we retrieved a list of all the candidates elected as Members of the European Parliament (MEP) and their corresponding countries, national parties and European coalitions:

https://www.wikidata.org/wiki/Wikidata:WikiProject_every_politician/European_Union/data/Parliament/Tenth

Further we added the MEPs who were not re-elected in 2024 and lost their seat at the European Parliament, retrieved from the list from this Wikipedia page:

https://en.wikipedia.org/wiki/List_of_MEPs_who_lost_their_seat_in_the_2024_European_Parliament_election

Language selection

We considered for our dataset the 24 official languages of the European Union: Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovene, Spanish, Swedish.

We added other popular languages in the European Union, namely we selected the languages spoken by more than 1% of the population⁵: Catalan, Russian, Turkish, Arabic, Galician, Chinese.

⁴ <https://www.wikidata.org/>

⁵ According to this table: https://en.wikipedia.org/wiki/Languages_of_the_European_Union#Knowledge. Retrieved from Wikipedia on January 23, 2025.

Finally, together with Catalan and Galician we also included Basque, as these are three official languages of Spain that were requested for being admitted as official languages of the European Union.⁶

Overall, the list includes 31 languages either official in the European Union (or requested for being admitted as such) or spoken by at least 1% of the European Union population.

Interaction data retrieval

The data was collected using the Wikipedia API⁷, a robust interface that provides programmatic access to Wikipedia's metadata, content, and revision histories. For each language, detailed revision histories were retrieved for both the main article and the associated talk page, when existing; a talk page associated with an article is a page used by the community to discuss any issue related to the collaborative creation of the article (Laniado et al, 2011).

Each revision includes comprehensive metadata, such as a unique revision ID ("revid"), parent revision ID ("parentid"), exact date and time when the revision was applied ("timestamp"), contributor username or IP address ("user"), contributors' ID ("userid"), and whether the revision was marked as minor ("minor"). Additional fields include the size of the page after the edit ("size"), the edit summary ("comment"), the parsed version of the comment ("parsedcomment"), and edit tags ("tags"). Data related to ORES scores ("orescores") estimating edit quality according to a machine learning model (Halfaker & Geiger, 2020), whether the edit was made by an anonymous user ("anon"), and the associated article title ("page_name") were also retrieved. For some revisions, specific fields may be hidden, such as comments ("commenthidden"), user information ("userhidden"), or entire edits ("suppressed"), depending on privacy or suppression policies.

The data extraction was designed to handle diverse page titles accurately by resolving interlanguage links and managing special characters. Systematic requests were made for the sets of pages and languages described before, and the resulting data was organized into structured language-specific files. Each file encompassed all revisions for articles and talk pages, with the article name and the country which is about (if there is such information), included as contextual attributes. This approach ensured efficient data retrieval, enabling a detailed exploration of multilingual editing dynamics on Wikipedia in the context of the 2024 European Parliament Elections.

⁶

<https://thediplotainSpain.com/en/2023/08/18/eu-receives-spains-request-to-recognise-catalan-basque-and-galician-as-official-languages/>

⁷ https://www.mediawiki.org/wiki/API:Main_page

Dataset description

In this section we first present some basic statistics about the dataset collected, and then report a few preliminary results of a high level analysis performed to provide a general description of the data by country, their temporal evolution, their distribution patterns and cross-language coverage.

Dataset general statistics

In Table 1, we report the basic statistics of the dataset collected. For each language edition included in the dataset, the table reports: the number of articles and of associated talk pages (when a talk page associated with an article exists and has at least one revision); the number of revisions made to the article (or number of edits) and to the talk page (or comments), and the sum of the two; the number of registered users who contributed at least an edit to an article in the corresponding language, the number of registered users who left at least a comment in a talk page, and the number of overall users who contributed (either articles or talk pages) in the corresponding language.

It should be noted that when we count users we only consider registered users, i.e. we exclude anonymous users, or users who edit without logging in, and are identified through their IP address. The rationale for excluding these users, in line with previous literature (Ortega Soto, 2012), is that they cannot be reliably identified and counted: the same users could appear from multiple IP addresses (or a same IP address could represent multiple users); furthermore, users who do not log in choose to edit anonymously and not to be identified by the community.

We observe that the number of articles varies from 1128 for the English Wikipedia, which by definition includes all the articles in the collection for how we built the dataset, and close to 1000 for Polish and French that include almost all the articles, to lower numbers for smaller language editions such as Irish, Latvian, Slovenian and Maltese.

Table 1: Basic statistics of the dataset collected, by Wikipedia language edition.

Language	Article pages	Talk pages	Article revisions	Talk revisions	Total revisions	Article users	Talk users	Total users
Arabic	276	238	10538	762	11300	486	25	487
Bulgarian	153	22	9841	229	10070	668	62	676
Catalan	340	136	21268	409	21677	1038	128	1056
Chinese	287	186	17079	761	17840	1203	116	1225
Croatian	92	16	4602	90	4692	449	38	453
Czech	297	59	24854	1391	26245	1500	208	1529
Danish	196	95	14496	897	15393	1132	164	1154
Dutch	354	109	47809	6227	54036	3050	520	3115
English	1128	1069	302847	40165	343012	21623	3649	22546
Estonian	165	65	6258	546	6804	423	68	427
Euskera	127	24	3990	66	4056	298	25	300
Finnish	668	59	27395	1554	28949	1529	222	1563
French	936	636	145879	15699	161578	7611	1304	7885
Galego	89	42	4294	113	4407	330	31	331
German	797	318	158792	52338	211130	11635	3180	12461
Greek	241	79	33256	2078	35334	1244	151	1264
Hungarian	223	103	19814	1091	20905	1188	182	1210
Irish	47	5	1469	16	1485	141	8	141
Italian	554	272	97264	7124	104388	4255	704	4397
Latvian	64	21	3428	93	3521	269	23	270
Lithuanian	96	15	5233	95	5328	416	33	420
Maltese	7	2	421	6	427	49	4	50
Polish	1010	116	64987	2102	67089	3596	267	3647
Portuguese	472	118	25490	833	26323	1866	155	1904
Romanian	188	137	12411	802	13213	952	111	975
Russian	370	194	29367	860	30227	2253	178	2277
Slovak	75	14	5819	111	5930	490	21	493
Slovenian	55	16	3170	43	3213	264	16	266
Spanish	415	190	82186	6149	88335	4769	753	5018
Swedish	310	110	38565	5505	44070	2507	457	2570
Turkish	217	196	10929	762	11691	758	77	771

The variables reported in the other columns are more or less aligned, and show that some small languages have little activity and a small base of active users editing on topics related to the European elections of 2024, while the most active communities both in terms of active users and number of edits and comments, apart from English, are German, French and Italian, followed by Spanish, Polish, Dutch and Swedish. The amount of activity seems to mirror a combination of different factors, where the number of active speakers is of course an important indicator but not the only one; the results, in line with previous literature, seem to suggest that other socio-economic factors also contribute, with higher presence in proportion to their population for nordic languages such as Finnish and Swedish.

Interestingly, Russian, Turkish and Chinese, which we included as relevant languages although not official languages of the European Union, have quite higher numbers of users and revisions than other official languages of the European Union. Russian, in particular, not only a popular language but also a language spoken by minorities in several Eastern European countries, results to have the second highest number of edits and users involved among all the Slavic and Eastern European languages included in the dataset, with only Polish having higher values.

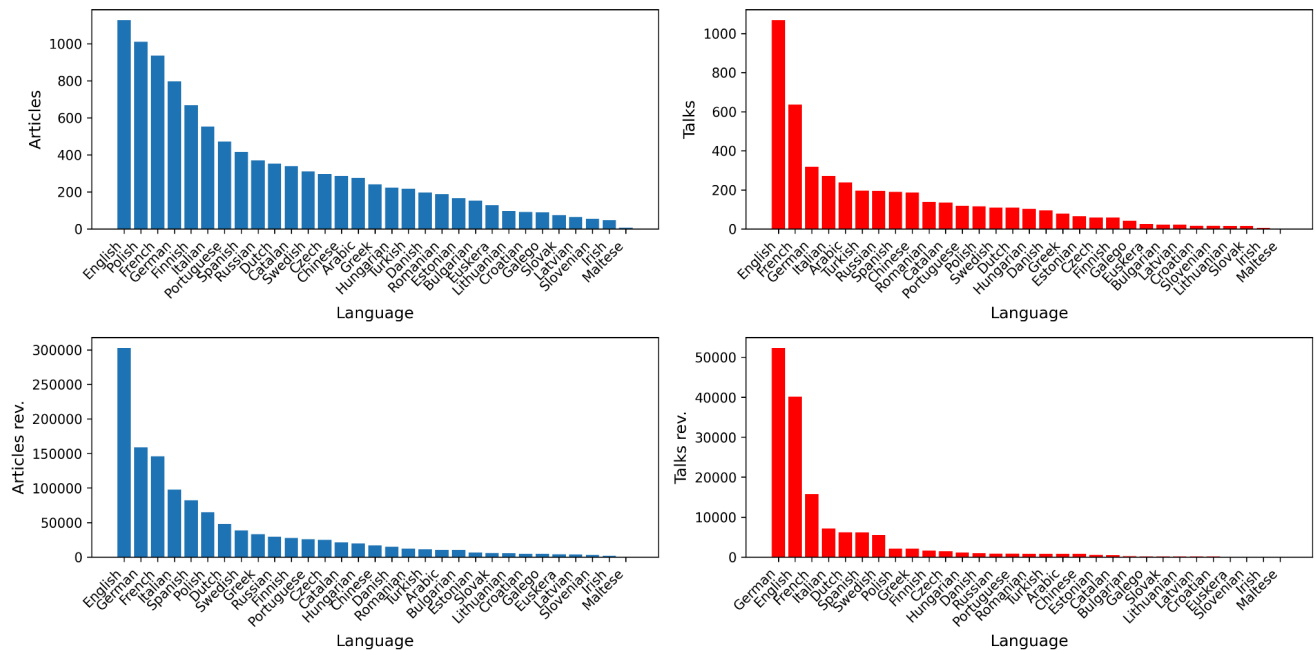


Figure 1: Ranking of the languages according to the number of Articles' (*left*) and Talks' (*right*) pages (*top*) and revisions (*bottom*).

To better appreciate these results, the bar charts in Figure 1 show the number of articles (top left, blue) and corresponding talk pages (top right, red) existing in each language edition, and the number of article (bottom left, blue) and talk page (bottom right, red) revisions per language edition. We observe that some languages like Polish or Finnish are among the top languages in terms of number of articles, but not in terms of activity levels; we can imagine that some editors created many articles in these languages, including articles on political actors from other European countries, but the community is not as large and active as for other language editions, such as the German, French or Italian that have the highest levels of activity behind English. Indeed, in terms of talk pages reviews, the German Wikipedia exhibits the highest level of activity even above English which is the common language of the European Union.

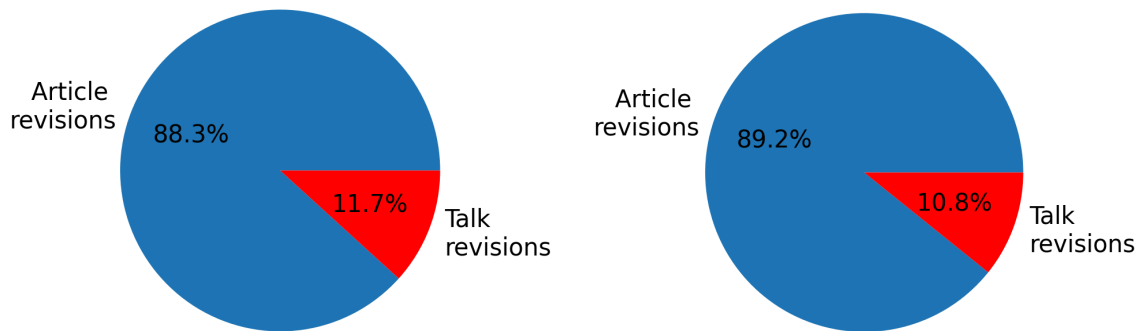


Figure 2: Pie chart reflecting the proportion of Article revisions and Talk pages revisions in all pages extracted from the English Wikipedia (*left*) and for all languages (*right*).

While the bar charts in Figure 1 are normalised to allow a comparison across languages with respect to article revisions and talk revisions separately, in Figure 2 we can appreciate the difference between these two spaces: the figure shows the proportion of revisions made to articles and to talk pages in the English Wikipedia (*left*) and in the whole dataset (*right*). In both cases talk pages receive about 11% of the activity, with a slightly higher proportion in the English Wikipedia.

Temporal evolution

Since each revision has associated a time stamp, we can now take a look at the temporal distribution of the activity. We start by observing the number of articles created over time until the

data extraction day (January 21st, 2025), in Figure 3 for the English Wikipedia: the bottom graph shows the number of pages created per day, among the ones considered in the dataset, while the top graph shows the corresponding cumulative number, i.e. the overall number of pages, which by definition is monotonically increasing. Figure 4 is analogous to Figure 3 but encompassing the whole dataset, considering articles in all languages together. In this case the cumulative graphs for articles and talk pages are shown in different graphs to improve their readability.

We observe significant peaks in the creation of new articles during the electoral campaign; interestingly, however, the higher peaks are present one-two months after the election; this suggests that many politicians running for the elections might not have a Wikipedia entry associated with them in English or in some other language, and such entries were created shortly after their election. On the other hand, the inset shows an increasing trend of new articles and talk pages during the campaign.

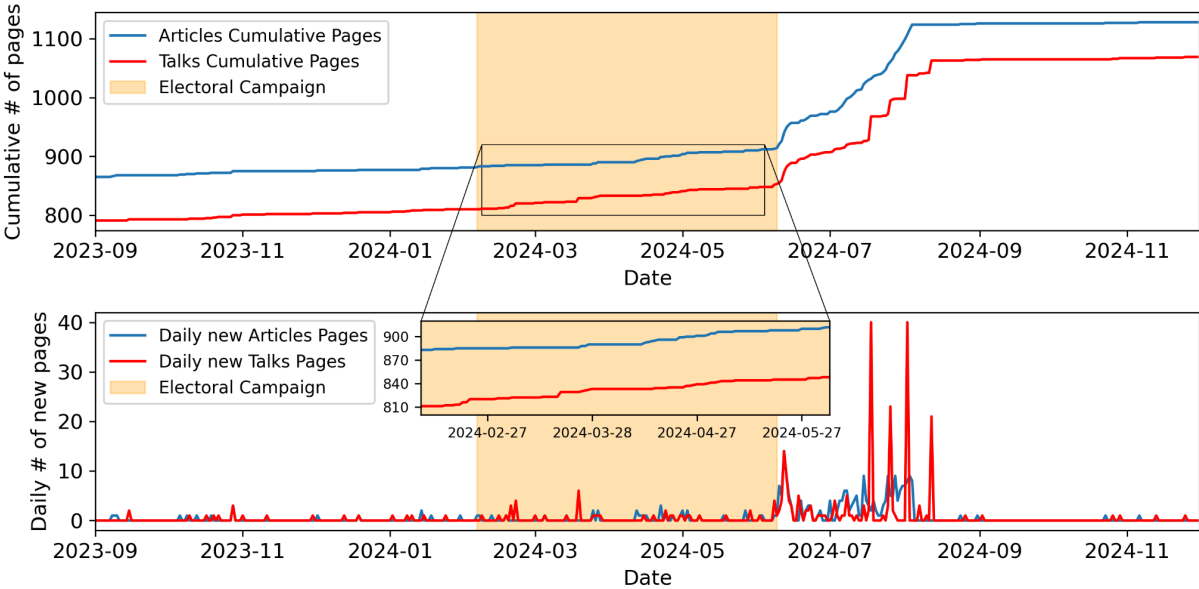


Figure 3: Cumulative number of articles' and talks' pages (upper) and daily number of new pages (lower) over time from September 1st, 2023 (2023-09-01) to December 1st, 2024 (2024-12-01) for all pages in English. The electoral campaign period, from February 6th to June 9th, is highlighted. The inset in the upper plot shows a zoom to the electoral campaign period to show its trend.

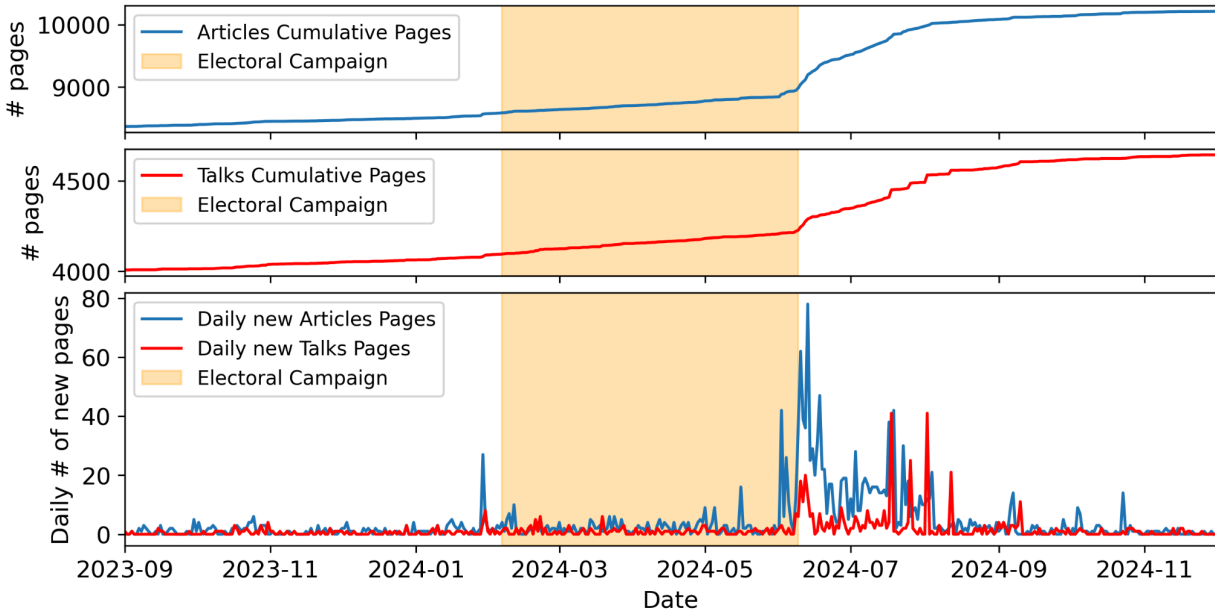


Figure 4: Cumulative number of articles (top) and talk pages (middle) and daily number of new pages (bottom) over time from September 1st, 2023 (2023-09-01) to December 1st, 2024 (2024-12-01) for all the language editions. The electoral campaign period, from February 6th to June 9th, is highlighted.

We then look at the number of revisions over time in the English Wikipedia (Figure 5) and in all language editions (Figure 6). In both cases we see the highest edits peak (blue line) right after the elections, while for talk pages revisions we find the highest peak about one month after the elections, around July 9th; this seems to be mostly a peak in some English Wikipedia talk pages, that is also quite visible when we aggregate all languages together. The peak has to do with the page “Patriots for Europe”, the group created by Orban and allied parties on July 8th, 2024.⁸

The red and blue lines, representing respectively article revisions and talk revision, show peaks that do not always overlap, as it was observed in previous literature (Kaltenbrunner & Laniado, 2012). In Figure 6, we observe higher activity during the electoral campaign, especially in terms of talk page revisions. This might be also due to restrictions to editing: pages about sensitive topics such as biographies of politicians running for elections may be protected or semi-protected, i.e. they can be edited only by registered editors, only by editors having a certain level of experience in the community, or only by administrators. This may explain higher levels of activity in talk pages for certain topics in certain languages; we will deepen into this, among many other aspects, in the analysis that we will perform on this dataset in task “T2.5 Computational analysis” and report in deliverable and report in deliverable D2.5 “Computational analysis report”.

⁸ See: <https://www.euronews.com/my-europe/2024/07/08/far-right-patriots-group-springs-to-third-force-in-european-parliament>

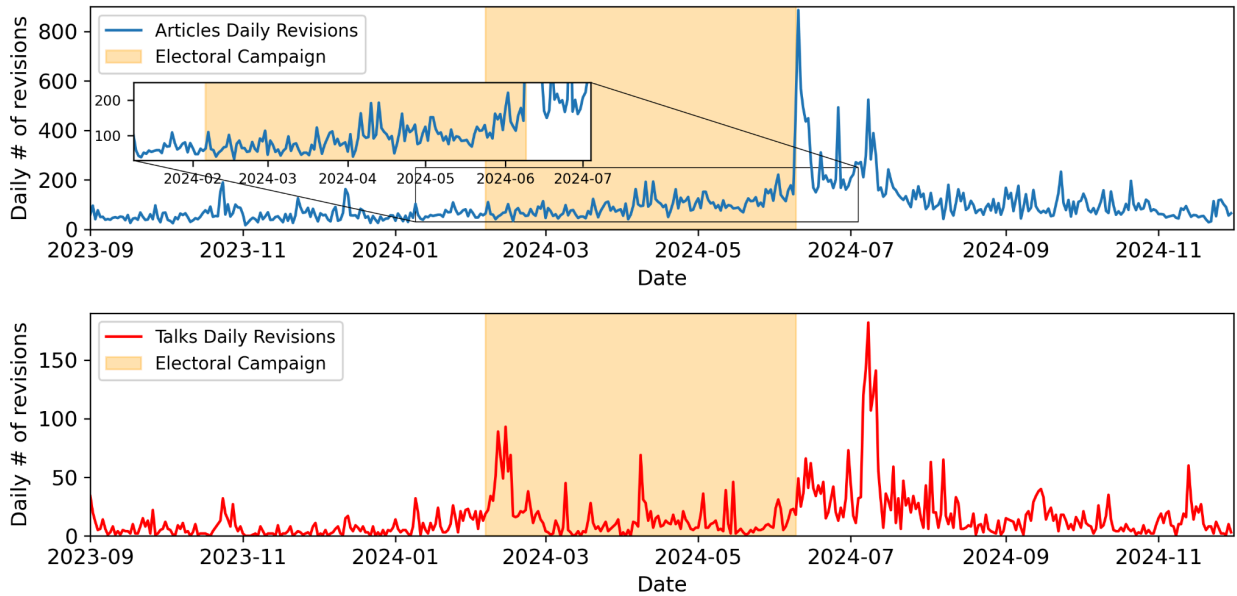


Figure 5: Number of daily revisions in articles (top, blue) and talk pages (bottom, red) from September 1st, 2023 (2023-09-01) to December 1st, 2024 (2024-12-01) for all pages in English. The electoral campaign period, from February 6th to June 9th, is highlighted. The inset in the upper plot shows a zoom to the electoral campaign period to show its trend.

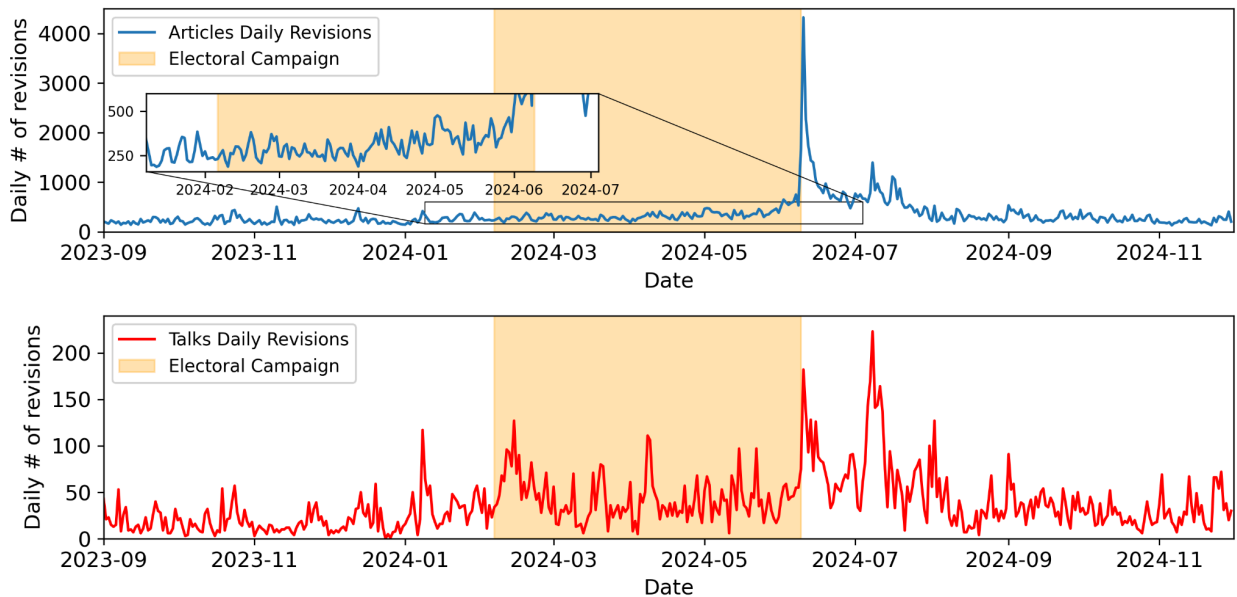


Figure 6: Number of daily revisions in articles (top, blue) and talk pages (bottom, red) from September 1st, 2023 (2023-09-01) to December 1st, 2024 (2024-12-01) for all pages in all languages. The electoral campaign period, from February 6th to June 9th, is highlighted. The inset in the upper plot shows a zoom to the electoral campaign period to show its trend.

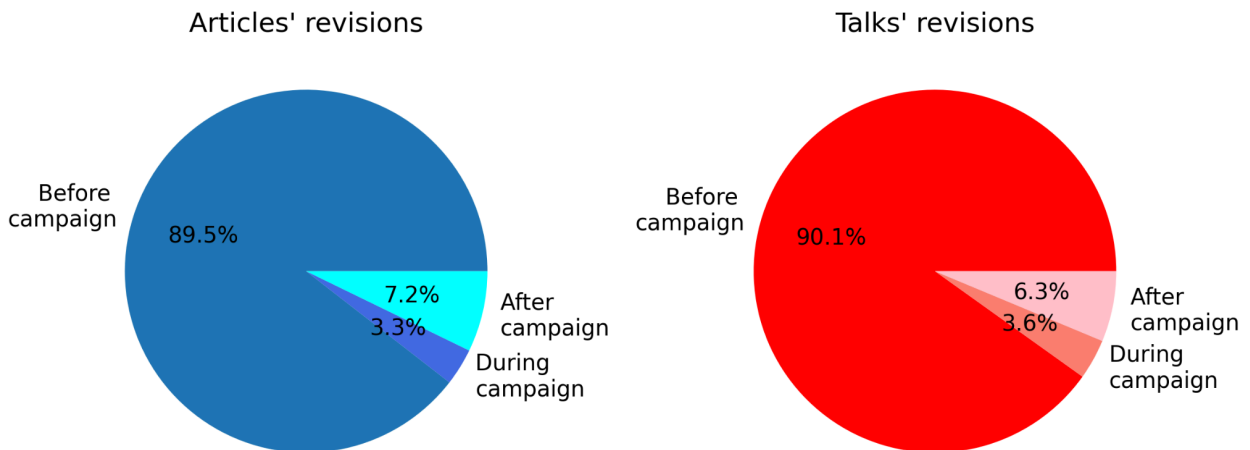


Figure 7: Pie chart reflecting the proportion of Articles' (left) and Talks' revisions (right) created before, during and after the Electoral campaign period, i.e. between February 6th and June 9th, 2024 from all language editions.

While in Figures 3, 4, 5 and 6 we only show data since September 2023 (for readability reasons and for allowing the reader to appreciate the trend during the electoral campaign and around the elections period), in Figure 7 we represent in a pie chart the proportion of revisions in articles (left) and talk pages (right) performed before, during and after the electoral campaign period over the total. As it can be noticed, even though many pages were created during the electoral campaign and afterwards (corresponding to the level of activity shown in Figures 5 and 6), it is interesting to notice that this still represents a small fraction of the activity in the overall history of the pages. Performing the computational analysis we will look more in detail at which articles, parties, topics, countries and languages were edited the most, both in absolute and relative terms, during the electoral campaign.

Activity distribution

In Figure 8, we show the distribution of activity on our selected pages in the English Wikipedia, by page (left) and by user (right). The plots in both cases show fat-tailed distributions typical of online social networks and online communities, and also in Wikipedia as shown by previous literature (Laniado et al, 2011; Tenorio-Fornés et al, 2022). A few pages attract most of the activity, while many less popular pages are characterised by low activity volumes; similarly, as we pass from the left plot to the right plot, we can observe that a few users are responsible for a high proportion of the activity (both in terms of article and talk revisions), while a large number of users only perform

a few edits. In the case of the distribution of revisions per user (right plot), the power law fit is even more straightforward than for the distribution of revisions per page (left plot). In this case, activity per user is well fitted by a power law with the same exponent ($\alpha \sim 2$) for both articles and talk pages.

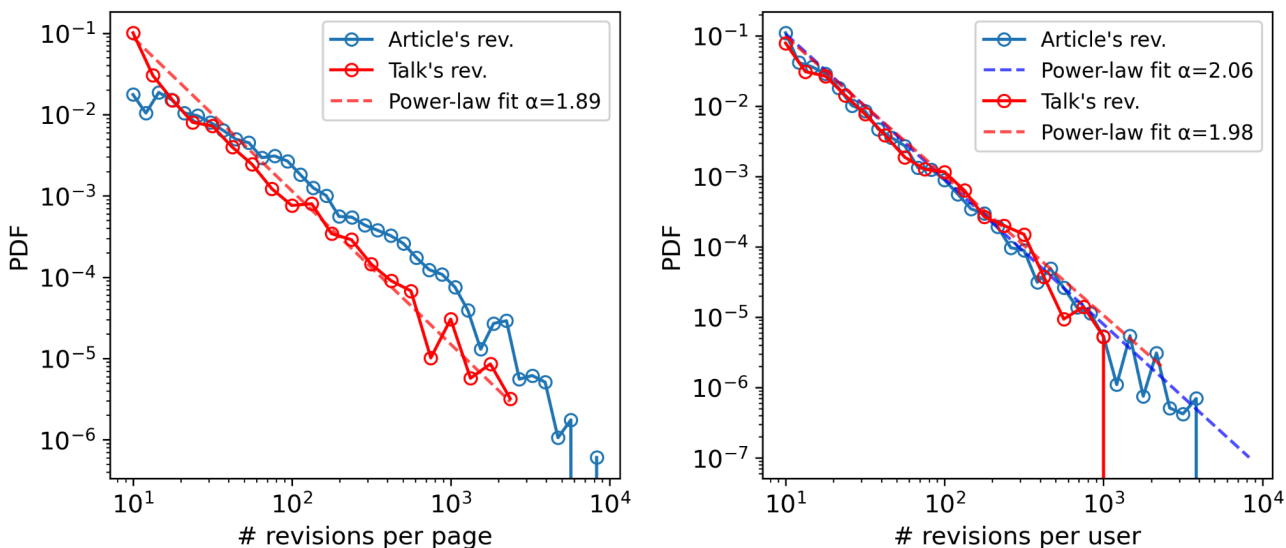


Figure 8: Distribution of the number of revisions per page (left) and revisions per user (right) for the English Wikipedia pages. Both distributions follow a fat-tail distribution, with an exponent determined by the fit, shown in the legend (when relevant).

In Figure 9, we observe analogous plots for all languages considered together; again we find fat-tailed distribution, well-fitted by a power law, in the case of the distribution of revisions per user, and with a similar trend for articles and talk pages. In the case of the distribution of the number of revisions per page (left plot), we observe a less clear pattern; in particular, the distribution is less skewed for pages having one or few revisions. This result suggests that, when a page exists, it easily has at least a few revisions. This might also be due in some cases to automatic tasks, such as templates added automatically by bots. On the other hand, articles with more than ten revisions follow the expected fat-tailed distribution, similar to talk revisions.

Of course it would be interesting to inspect language specific patterns, but this would take a lot of space given the number of languages, and would go beyond the scope of this document; language specific analysis will be performed in the next task (T2.5 Computational analysis) and reported in deliverable D2.5.

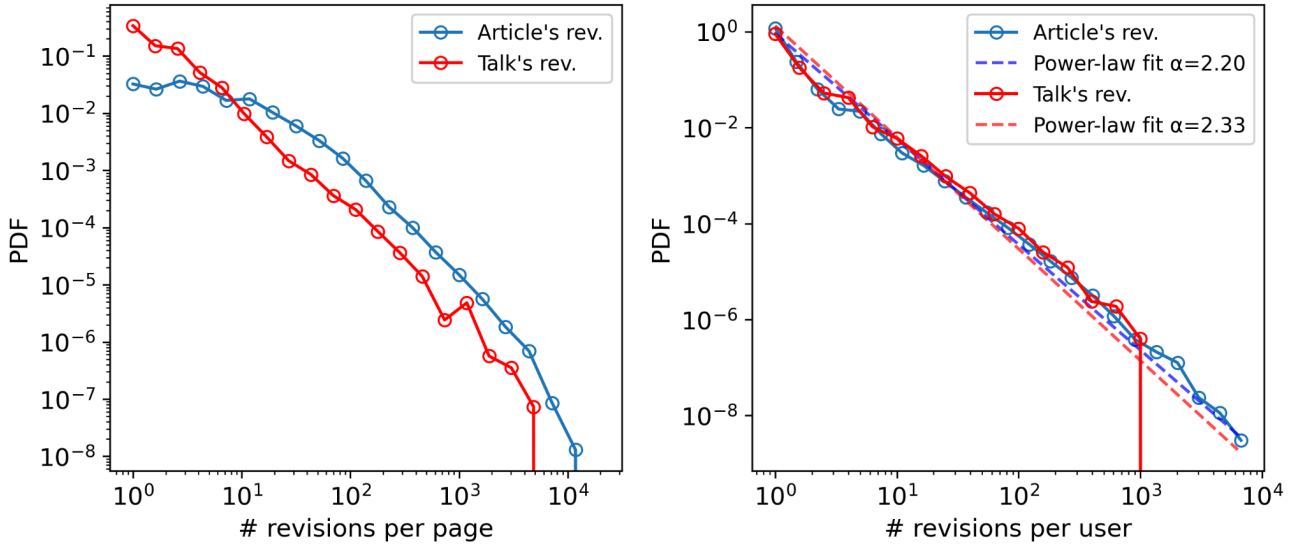


Figure 9: Distribution of the number of revisions per page (left) and revisions per user (right) for all pages in all languages considered (full dataset). Both distributions follow a fat-tail distribution, with an exponent determined by the fit, shown in the legend (when relevant).

Cross-language coverage

Finally, we take a look at how each language covers the content related to each country. Since our pages are related to European elections, all Wikipedia articles about politicians and their regional parties have a country of origin for which we can label the Wikipedia article. The country associated with a Wikipedia page is the same for all languages.

Intuitively we can expect each language to cover mainly the content related to the territories where it is spoken, as also documented by previous literature (Miquel-Ribé & Laniado, 2021); the question we ask is to what extent this happens, to what extent each language covers political actors of other countries, and which are preferential relationships between languages and countries.

In Figure 10, we represent a matrix of languages and countries, following the approach in (Miquel-Ribé & Laniado, 2018). Each matrix element shows which proportion of the content related to each country, in terms of number of articles, is covered by each language in our dataset. The order of languages and countries is such that it makes it easier to identify the most relevant languages that cover each country. Darker tones of red represent higher proportions of coverage while whiter represent a lower proportion. We observe a strong pseudo-diagonal of darker cells, where each language crosses its corresponding countries. For example at the intersection of

German with Germany and Austria, of Dutch with Netherlands and Belgium, or of Greek with Greece and Cyprus. The English language covers entirely all countries as it was used as the basis for building the dataset. We can see that Polish tends to also cover most of the articles related to all European countries.

Moreover, we observe some connections, such as French having a 99% coverage of content related to Germany, or Finnish covering 100% of content on Sweden, Denmark and Estonia, and Swedish in turn covering 96% of content on Finland.

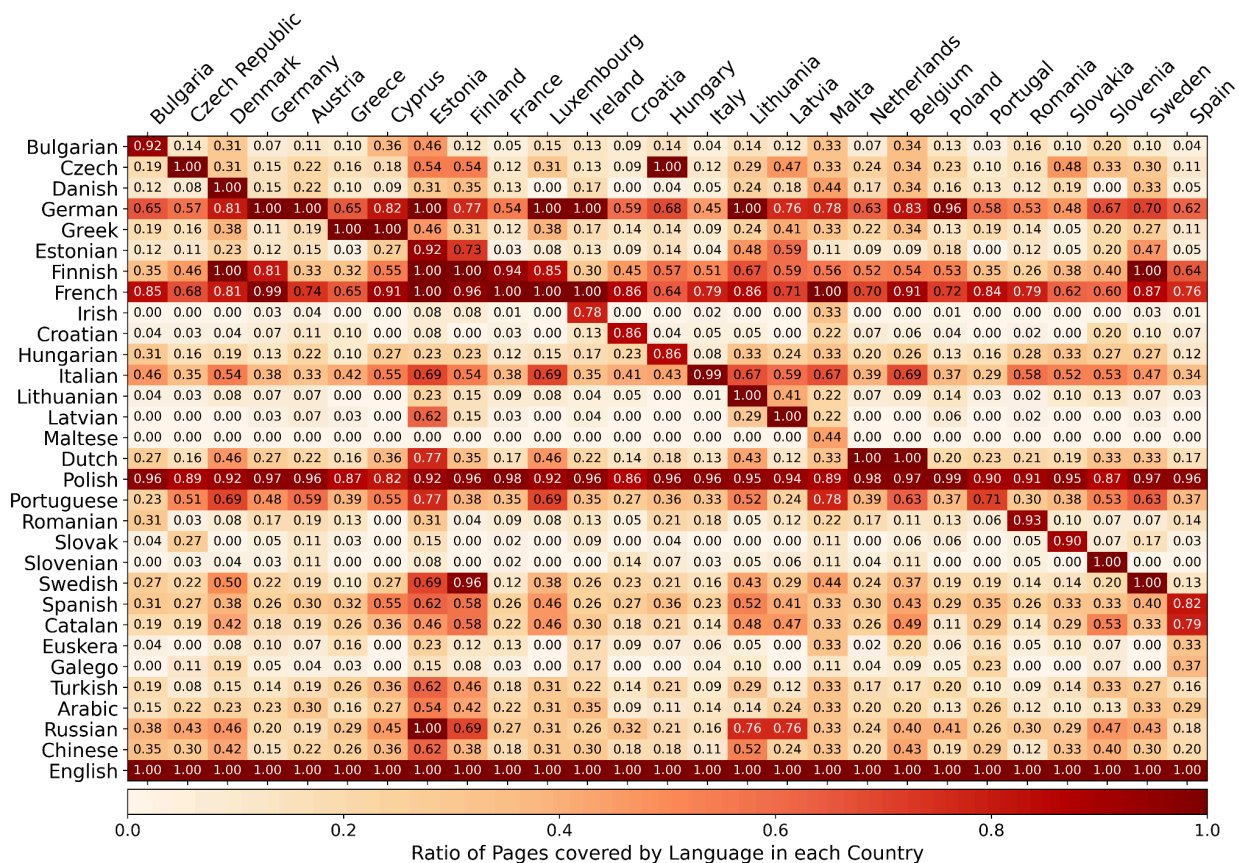


Figure 10: Ratio of pages about each of the countries that is written in a certain language. The number of pages in a given language about a certain country is normalized by the total number of pages about that country, which always corresponds with the number of pages in English, according to our data extraction methodology. Countries are ordered to approximately match the expected language. The color indicates the value of the ratio, as described in the colorbar.

While in Figure 10 we only consider the number of pages existing in a given language about a given country, now we give one step ahead and we consider the amount of activity and users in each linguistic community about each country. In Figure 11 we show the fraction of total revisions

(edits to articles and talk pages) and fraction of distinct registered users participating, i.e. making at least one edit to an article or talk page about a given country.

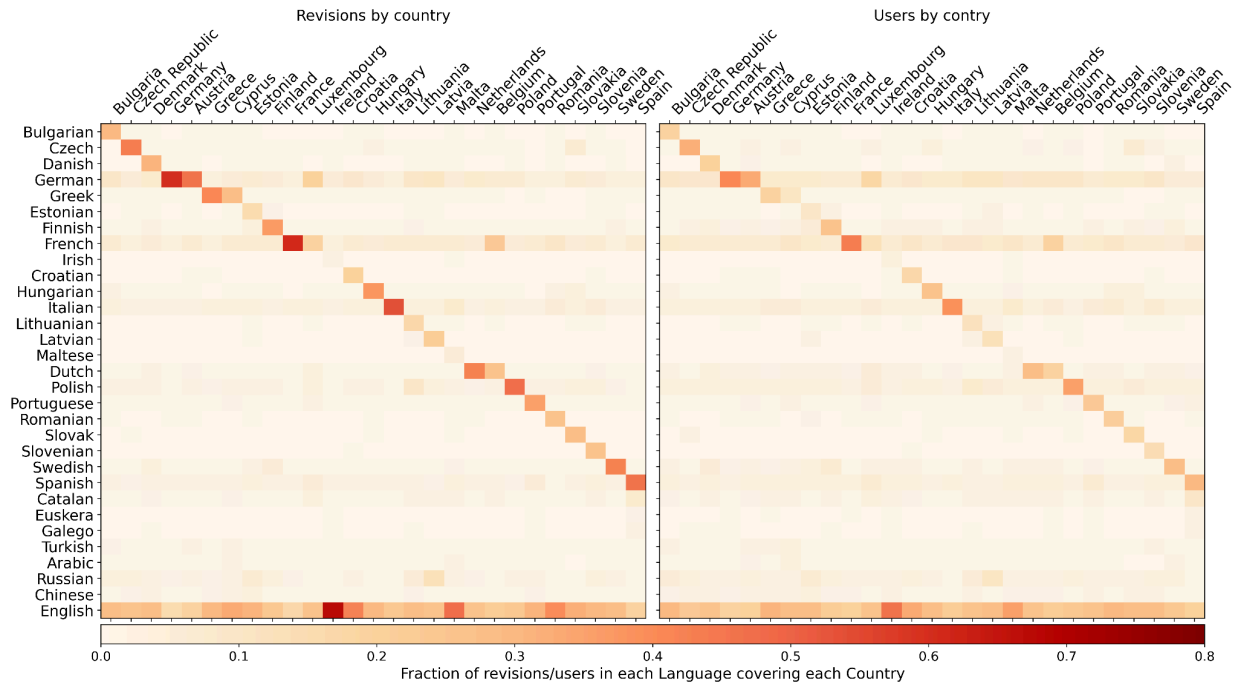


Figure 11: Fraction of revisions (left) and users (right) in each language about each country. The number of revisions/registered users (from Articles and Talks) writing in a given language about a certain country is normalized by the total number of revisions/registered users. Countries are ordered to approximately match the expected language. The color indicates the value of the fraction, as described in the colorbar.

Here we can appreciate that the pseudo-diagonal is even more marked in both graphs, as the activity of each linguistic community tends to concentrate mostly on content related to the associated countries and territories (Miquel-Ribé et al, 2021).

Conclusions

In this document we have documented the work in task T2.2 - Collection of digital traces, which constitutes the basis for performing the computational analysis.

The deliverable describes the data collection process and some basic statistics about the dataset collected. We have also shown a few preliminary high level analyses of some basic aspects of the dataset, such as temporal evolution, activity distribution and cross-language coverage.

The dataset was built with the intention of following a rigorous and replicable procedure, minimising the possibility of bias and arbitrary choices, for collecting the communities' interactions around content that is clearly related to the electoral campaign. At the same time, we remain open to adding more pages to our dataset, including relevant pages that we may have missed with our collection procedure, and possibly relevant internal community pages that we could also receive as input from the communities.

The statistics reported show that for some small languages there is so little data that it may be not meaningful to be analysed. For the next steps of the computational analysis, we plan to set different thresholds according to different kinds of methods used, to ensure a minimum level of activity that makes the analyses meaningful.

The preliminary high level results presented in this deliverable already show some interesting directions for the computational analysis, such as delving into the topics objects of activity peaks in articles and talk pages to identify and inspect controversies, studying activity distribution and community dynamics in different language editions, investigating the attention devoted by each language community to the different European political parties, or considering the sources cited in the articles and in their different revisions as done by Baigutanova (2023) or D'Ignazi et al (2024).

References

Baigutanova, A., Saez-Trumper, D., Redi, M., Cha, M., & Aragón, P. (2023). A Comparative Study of Reference Reliability in Multiple Language Editions of Wikipedia. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 3743-3747).

D'Ignazi, J., Kaltenbrunner, A., Mejova, Y., Tizzani, M., Kalimeri, K., Beiró, M., & Aragón, P. (2024). Language-Agnostic Modeling of Source Reliability on Wikipedia. *arXiv preprint arXiv:2410.18803*.

Kaltenbrunner, A., & Laniado, D. (2012, August). There is no deadline: time evolution of Wikipedia discussions. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (pp. 1-10).

Halfaker, A., & Geiger, R. S. (2020). Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4 (CSCW2), 1-37.

Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When the Wikipedians talk: Network and tree structure of Wikipedia discussion pages. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 5, No. 1, pp. 177-184).

Laniado, D., Mauri, M., & Borra, E. (2024). Chapter 6. Exploring the evolution of Wikipedia articles through Contropedia. In *Investigating Wikipedia: Linguistic corpus building, exploration and analysis* (pp. 156-177). John Benjamins Publishing Company.

Miquel-Ribé, M., & Laniado, D. (2018). Wikipedia culture gap: quantifying content imbalances across 40 language editions. *Frontiers in physics*, 6, 54.

Miquel-Ribé, M., & Laniado, D. (2019). Wikipedia cultural diversity dataset: A complete cartography for 300 language editions. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 13, pp. 620-629).

Miquel-Ribé, M., & Laniado, D. (2021). The Wikipedia Diversity Observatory: helping communities to bridge content gaps through interactive interfaces. *Journal of internet services and applications*, 12(1), 10.

Miquel-Ribé, M., Laniado, D., & Kaltenbrunner, A. (2021). The role of local content in Wikipedia: A study on reader and editor engagement. *Área Abierta*. 2021; 21 (2): 123-151.

McDowell, Z. J. (2024). Wikipedia and AI: Access, representation, and advocacy in the age of large language models. *Convergence*, 30(2), 751-767.

Ortega Soto, J. F. (2012). Wikipedia: A quantitative analysis. Doctoral dissertation. <https://burjcdigital.urjc.es/bitstream/handle/10115/11239/thesis-jfelipe.pdf>

Saez-Trumper, D. (2019). Online disinformation and the role of wikipedia. *arXiv preprint arXiv:1910.12596*.

Smith, C. E., Yu, B., Srivastava, A., Halfaker, A., Terveen, L., & Zhu, H. (2020). Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

Tenorio-Fornés, Á., Arroyo, J., & Hassan, S. (2022). Participation in wiki communities: reconsidering their statistical characterization. *PeerJ Computer Science*, 8, e792.

Vincent, N., & Hecht, B. (2021). A deeper investigation of the importance of Wikipedia links to search engine results. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1-15.

January 2025

Disclaimer

The sole responsibility for any content supported by the European Media and Information Fund lies with the author(s) and it may not necessarily reflect the positions of the EMIF and the Fund Partners, the Calouste Gulbenkian Foundation and the European University Institute.

<https://gulbenkian.pt/emifund/disclaimer/>