WEB2WIKI: Characterizing Wikipedia Linking Across the Web

Veniamin Veselovsky[1], **Tiziano Piccardi**[2], **Ashton Anderson**[3], **Robert West**[4], **Akhil Arora**[5]
[1] Princeton University
[2] Johns Hopkins University
[3] University of Toronto
[4] EPFL
[5] Aarhus University
veniamin@princeton.edu, piccardi@jhu.edu, ashton@cs.toronto.edu, robert.west@epfl.ch, akhil.arora@cs.au.dk

Abstract

Wikipedia is one of the most visited websites globally, yet its role beyond its own platform remains largely unexplored. In this paper, we present the first large-scale analysis of how Wikipedia is referenced across the Web. Using a dataset from Common Crawl, we identify over 90 million Wikipedia links spanning 1.68% of Web domains and examine their distribution, context, and function. Our analysis of English Wikipedia reveals four key findings: (1) The topics of Wikipedia articles referenced on the Web differ from those cited on Reddit and those prominent within Wikipedia's own link structure, (2) Wikipedia is most frequently cited by news and science websites for informational purposes, while commercial websites reference it less often. (3) The majority of Wikipedia links appear within the main content rather than in boilerplate or user-generated sections, highlighting their role in structured knowledge presentation. (4) Most links (95%) serve as explanatory references rather than as evidence or attribution, reinforcing Wikipedia's function as a background knowledge provider. While this study focuses on English Wikipedia, our publicly released WEB2WIKI dataset includes links from multiple language editions, supporting future research on Wikipedia's global influence on the Web.

1 Introduction

As the de facto encyclopedia of the Internet, Wikipedia is one of the most visited websites globally (Singer et al. 2017) and is widely cited as an authoritative source of information (Mesgari et al. 2015). This unique status and prosperity have further inspired rich bodies of academic literature on Wikipedia, providing insights about its structure (Holloway, Bozicevic, and Börner 2007; Arora, West, and Gerlach 2024), readership (Lemmerich et al. 2019), knowledge gaps (Zia et al. 2019), as well as user behaviors (Arora et al. 2022; Piccardi et al. 2023; Piccardi, Gerlach, and West 2022; Piccardi 2022; Arora 2024). Initial research has also examined Wikipedia's role beyond its own platform, analyzing its impact on law and science (Thompson et al. 2022; Thompson and Hanley 2018), its importance on social media (Vincent, Johnson, and Hecht 2018), and its influence on search engine quality (McMahon, Johnson, and Hecht 2017; Vincent et al. 2019). However, the extent of Wikipedia's presence across the broader Web remains unexplored, leaving a significant gap in our understanding of its external influence. While most of its traffic originates from search engines, a substantial portion comes via direct links to its articles (Piccardi et al. 2023). Platforms such as YouTube and news websites integrate Wikipedia links to provide additional context for specific content or controversial topics[1]. Similarly, users of platforms like Quora or Reddit frequently reference Wikipedia articles to support their arguments (Moyer et al. 2015).

Given its central role, Wikipedia should not be studied in isolation; mapping its role across the Web is critical. First, understanding how Wikipedia-generated knowledge is used beyond its own platform can reveal key insights. These include which articles are most frequently referenced, the contexts in which they appear, the types of websites linking to Wikipedia, and the motivations behind these links.

Second, it can help quantify the extent and size of Wikipedia by measuring its reach, which is critical in assessing its value and understanding its global influence. These insights can inform policies by assessing Wikipedia's potential impact on public discourse, indicating areas where Wikipedia could play a central role in shaping public understanding. Finally, it can open up a new paradigm for studying what matters on Wikipedia through the rich world of public links. It reveals which topics are disproportionately referenced externally relative to their prominence within Wikipedia, and informs how Wikipedia content is used across the broader Web.

Specifically, we formulate three research questions:

- **RQ1:** *What* types of Wikipedia articles are referenced most on the Web?
- **RQ2:** *Where* are Wikipedia articles most likely to be linked?
- **RQ3:** *Why* do people reference Wikipedia articles on external websites?

To answer these research questions, we generated a dataset called WEB2WIKI containing all Wikipedia links from Common Crawl—the largest public dump of HTML content on the Web, and conducted a large-scale analysis of how English Wikipedia is linked across the Web. We begin by examining which types of Wikipedia articles are disproportionately linked across different parts of the Web. To contextualize these patterns and provide a complete picture, we

[1] https://support.google.com/youtube/answer/7630512

compare how frequently articles are linked on the broader Web and the social Web, with a specific focus on Reddit. Following prior work that uses in-degree as a measure of Wikipedia article importance (Fortunato et al. 2006; Thalhammer and Rettinger 2016), we use the number of incoming links from Wikipedia's internal network as a baseline proxy for an article's significance.

Then, to examine where Wikipedia is linked, we conduct (1) a domain-level analysis of linking websites and (2) an in-page segmentation of Wikipedia references. Our findings support the expectation of Wikipedia as an authority for validating knowledge: Wikipedia articles are far more likely to be referenced on News, Science, and Society websites when compared to Business and Shopping websites. When Business and Shopping websites do link to Wikipedia, they are more likely to place links in the boilerplate portion of the page rather than within the main content.

Finally, to understand why Wikipedia is referenced, we identified two primary motivations through iterative coding of Wikipedia links across the Web. The predominant reasons are "evidence" and "delegation'.' Evidence-based linking occurs when a website embeds Wikipedia-sourced content verbatim or cites Wikipedia to support a claim, fact, or media resource. In contrast, delegation-based linking is broader in scope. Websites use Wikipedia for contextual explanations (e.g., linking to the Wikipedia article on HTTP cookies when explaining privacy policies), for background information on niche topics, or as an integrated reference across various content types.

Alongside these findings, we release the WEB2WIKI dataset to facilitate further research on this topic, advance our understanding of Wikipedia's role on the Web, and serve as a foundation for future analyses of linking structures across the Web.

2 Related Work

The Wikimedia Foundation has long emphasized the importance of studying Wikipedia's broader connections to the Web. In 2015, its "New Research Directions" identified this as a key priority (Taraborelli 2015). The relative research roadmap highlights the need to determine what types of knowledge Wikimedia must acquire to better serve its role on the Web (WMF 2022). This work supports that agenda by offering a descriptive analysis and releasing the first dataset that maps Wikipedia's position and usage across the Web.

Value to the Web. Most prior studies have focused on specific platforms, particularly Wikipedia's role in social media and search engines. A theme in this line of work has been quantifying the value that Wikipedia offers these sites. Social media platforms, particularly Reddit and StackOverflow, benefit significantly from Wikipedia references, whereas Wikipedia itself receives little direct traffic in return (Vincent, Johnson, and Hecht 2018). This phenomenon has been described as the "paradox of reuse": as Wikipedia increasingly powers external technologies, users may rely on its content without ever visiting the site directly (McMahon, Johnson, and Hecht 2017). Other studies focused on the role Wikipedia plays on search engines, finding that Wikipedia's content provides a large benefit for Google being omnipresent in search results and a critical tool for effective information retrieval due to the decrease in search result quality with Wikipedia removal (Vincent et al. 2019; McMahon, Johnson, and Hecht 2017). Extending this research beyond Reddit, StackOverflow, and search engines, a recent dataset release contains all mentions of Wikipedia articles on Twitter, providing new opportunities for analysis (Meier 2022).

Additionally, Wikipedia articles being invoked in the Reddit r/TodayILearned (TIL) subreddit leads to a nontrivial increase in Wikipedia viewership (Moyer et al. 2015), and Wikipedia is a gateway to the Web, with large swathes of Wikipedia browsing activity leading to other websites via external links present in Wikipedia articles (Forte et al. 2018; Piccardi et al. 2021).

Finally, Wikipedia's value to the Web has been studied through the lens of plagiarism detection (Alshomary et al. 2019). They construct a dataset of Wikipedia references extracted from Common Crawl to identify unattributed content on the Web, and then detect plagiarism by using hashing to find matching pairs of textual content between the Web and Wikipedia.

Societal impact beyond the Web. A limited body of research has explored Wikipedia's broader societal impact. One study conducted a randomized field experiment on legal articles in Ireland and found that judges heavily rely on Wikipedia in shaping judicial behaviour (Thompson et al. 2022). Another work similarly showed that Wikipedia articles are used as review pages for many scientific analyses (Thompson and Hanley 2018). Further studies have linked Wikipedia to measurable economic effects. For example, Wikipedia content influences tourism revenue, as destinations with better Wikipedia coverage attract more visitors (Hinnosaar et al. 2023) and the value of traffic Wikipedia drives to external websites–if it were acquired through online advertising–would amount to millions of dollars (Piccardi et al. 2021).

3 WEB2WIKI: Data and Methodology

Data and code. The dataset used in this paper was extracted from the February 2021 Common Crawl dump[2]—the largest public scrape of HTML pages on the Web created prior to February 2021—through a regex search of `<a>` tags, the HTML element for hyperlinking. Specifically, the WEB2WIKI dataset, released with this article, contains a large sample of the publicly accessible HTML pages that link to Wikimedia projects, as well as the bare webpage–article link pairs. Given the web-scraping policies of Common Crawl (cf. Appendix. A), the dataset contains links to Wikipedia articles available on the surface Web—the portion of the Web accessible to the general public and searchable with standard Web search engines[3]—, excluding websites

[2]https://commoncrawl.org/2021/03/february-march-2021-crawl-archive-now-available/

[3]https://en.wikipedia.org/wiki/Surface_web

Language	#Domains	#Webpages with links to Wikipedia	#Wikipedia articles linked on the Web	#Incoming Web-links	%Wikipedia articles responsible for 80% of the Web-links
English (EN)	940,239	14,462,267	3,619,416	48,829,702	9%
German (DE)	155,887	2,128,814	824,981	5,882,838	17%
French (FR)	96,875	1,465,997	692,220	4,925,889	17%
Japanese (JA)	56,651	818,110	656,086	4,455,828	20%
Spanish (ES)	88,903	1,262,583	486,811	3,629,491	17%
Russian (RU)	53,219	1,105,698	465,283	2,705,021	19%
Italian (IT)	55,329	851,533	350,046	2,463,232	16%
Portuguese (PT)	27,350	413,034	226,713	1,194,040	22%
Swedish (SV)	14,716	256,776	221,530	1,138,989	23%
Dutch (NL)	33,953	366,933	206,727	892,750	29%
Vietnamese (VI)	11,190	146,623	101,427	882,410	12%
Polish (PL)	23,012	376,167	161,395	752,888	23%

Table 1: Summary statistics of the WEB2WIKI dataset, portraying the number of distinct domains and webpages that link to Wikipedia, the number of distinct Wikipedia articles that are linked on the Web, the total number of incoming Web-links to Wikipedia, and the percentage of Wikipedia articles that are responsible for 80% of the Web-links, respectively, across the 12 most linked language versions of Wikipedia on the Web.

such as Facebook, Reddit, Twitter, Quora, or YouTube that require registering an account to see their content.

All code required to reproduce the analyses presented in this paper is available at https://github.com/vminvsky/web2wiki. The dataset will be made publicly available on Zenodo (cf. GitHub repository for details) under an open license (CC-BY-SA 4.0).

Data pre-processing and summary

To conduct the analyses presented in this paper, the first step involves processing each webpage to retrieve all Wikipedia links found within an `<a>` tag. Note that by restricting our focus solely to links that point to Wikipedia in English, we exclude the links to all other Wikimedia projects (*Wikiquote, Wikinews, Wikiversity, Wiktionary, Wikisource, Wikibooks, Wikivoyage*) from our analyses. However, these links are included in the full WEB2WIKI dataset. Starting from a dataset of 2.7 billion web pages—amounting to 280 TiB of uncompressed content—our pre-processed dataset contains 90,805,367 links to Wikipedia articles, spanning 1.68% of the domains in the Common Crawl dump.

Diving deeper into the dataset statistics (Table 1), English-language Wikipedia stands out for several reasons. First, it dominates the WEB2WIKI dataset, accounting for over 50% of all links, meaning it is referenced almost 10 times more frequently than the next most-used language (German). Second, links to English Wikipedia are highly concentrated: 80% of incoming links target just 9% of English articles—a significantly more skewed distribution than observed in other languages. Third, a large proportion (58%) of English Wikipedia articles are linked on the broader Web, compared to German (33%), French (30%), Spanish (30%), and Dutch (10%). Given these factors–along with the authors' proficiency in English–we focus our analysis on English Wikipedia, which corresponds to 48,829,702 links from 14,462,267 distinct webpages across 940,239 domains.

Overall, the distribution of Wikipedia links per webpage follows a long-tail pattern, with an average of 3.4 links per page, while some pages contain over 4,000 Wikipedia links.

Methodology

We focus on three important questions (*what*, *where*, and *why*) to ground our understanding of the true extent of Wikipedia's presence on the Web.

Measuring article importance. We first address which Wikipedia articles are most salient on the Web by comparing their presence in general web links to their use on the social Web, specifically in Reddit posts–a platform not included in Common Crawl data. Our data consists of all comments and submissions on Reddit from its inception in December 2005 to June 2022 (Baumgartner et al. 2020). This includes almost 13 billion comments by 89 million users and 5 billion submissions by 48 million users. To contextualize these patterns, we use Wikipedia's in-degree as a baseline proxy for article importance within the platform. We compute in-degrees from the Wikipedia XML dump released concurrently with the Common Crawl data, considering only internal links present in the body of each article (Mitrevski, Piccardi, and West 2020). This measure serves as a useful reference point, as it implicitly reflects an article's significance in Wikipedia's internal link network (Fortunato et al. 2006; Thalhammer and Rettinger 2016).

We then analyze which "types" of Wikipedia articles are disproportionately referenced on both the broader Web and Reddit.

ORES topics. To obtain a broader classification of Wikipedia articles that are highly represented on the Web, we use Wikipedia's ORES topic model (Halfaker and Geiger 2020). ORES classifies articles by topic and clusters them into sixty-four categories[4], grouping articles into broad themes such as *computing*, *biology*, and *society*.

[4]https://www.mediawiki.org/wiki/ORES/Articletopic

Websites topics. Next, we examine where Wikipedia articles are linked by analyzing the topics of websites that reference Wikipedia. To classify webpage topics, we use Homepage2Vec (Lugeon, Piccardi, and West 2022), a model that predicts a webpage's topic based on its content. This classification follows Curlie's taxonomy[5], which defines 14 distinct topics, including *News*, *Science*, and *Shopping*.

Webpage segmentation. We then analyze where *within* a webpage, the Wikipedia article link appears. To achieve this, we manually define a set of structural rules to segment webpages into three distinct sections: boilerplate, main content, and user contributions. To develop these segmentation rules, the authors manually inspected 500 random webpages containing Wikipedia links.

Boilerplate includes static elements that remain consistent across different pages, such as headers and footers. The main content encompasses the primary text of a webpage, where most of the substantive information is presented. Finally, user contributions capture Wikipedia links shared through interactive user-generated content, such as those appearing in comment sections and discussion threads.

4 What Articles Are Linked

The first question we examine is what types of Wikipedia articles are linked across the Web. A straightforward approach would be to count the number of distinct domains linking to a given article. However, interpreting these counts requires a meaningful baseline. As introduced before, we use Wikipedia in degrees to contextualize web links as a reference measure. This baseline assumes that an article is important if it is heavily referenced within Wikipedia itself. By comparing Wikipedia links on the general Web and Reddit to in-links within Wikipedia, we quantify how external references differ from Wikipedia's internal linking structure–comparing semantic connections within Wikipedia with how articles are referenced externally on the broader Web and the social Web (Reddit).

The Pearson correlation between external website links and Wikipedia article in-degree (number of incoming links) shows a significant positive relationship ($r = 0.5$, $p < 0.001$), indicating that articles with high in-degree (i.e., those referred by many other concepts) are more likely to be linked from external websites. By comparing how the three groups–the Web, Reddit, and Wikipedia itself–link to distinct topics, we identify both alignments and divergences in their referencing patterns.

Wikipedia-internal links, external web references, and Reddit citations each have distinct distributions over ORES topics. To quantify these distributions, we obtained the probability distribution–over the 64 topics–of Wikipedia articles linked from each source.

We define the *Wikipedia-internal probability* of a topic k as the likelihood that a randomly selected Wikipedia internal link points to an article on that topic. Similarly, we define the *Web probability* as the probability that an external website links to a Wikipedia article in a given ORES topic,

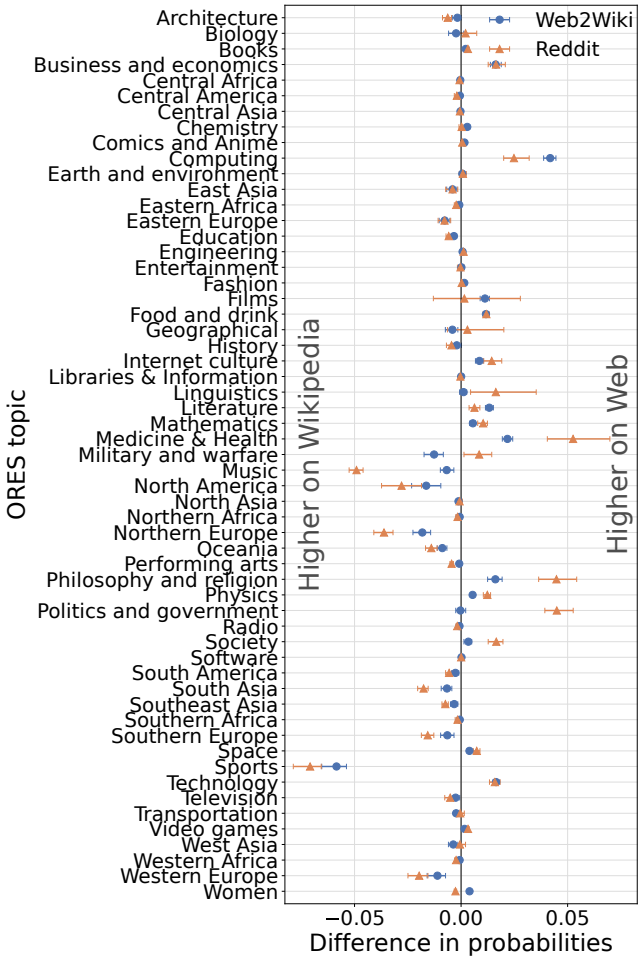

Figure 1: Comparison of proportion of Wikipedia-internal links, Web links, and Reddit shares of Wikipedia. A higher score means proportionally more in-links from external websites than as Wikipedia-internal links. Error bars represent 95% bootstrapped confidence intervals.

and the *Reddit probability* as the probability that a Wikipedia article cited on Reddit belongs to a specific ORES topic. By comparing these three probability distributions, we identify which topics are disproportionately referenced on the Web and Reddit relative to Wikipedia's internal linking structure. This analysis reveals differences in how various platforms emphasize certain topics, shedding light on external vs. internal knowledge representation. An important distinction is that the Wikipedia-internal probability is computed at the link level, whereas the Web probability is computed at the domain level. This reflects differences in the underlying data: within Wikipedia, multiple links from a single article contribute to its internal structure, while on the Web we aggregate at the domain level to avoid overrepresenting sites that embed many repeated links to the same topic. The Reddit probability, computed at the comment level, is structurally closer to the Wikipedia-internal measure. We therefore interpret these distributions as com-

[5]https://curlie.org/

plementary perspectives on topical emphasis, and focus on relative patterns rather than strict one-to-one comparisons.

Figure 1 illustrates the differences in topic distributions across Wikipedia-internal links, Web links, and Reddit shares. The x-axis represents the probability difference for a given ORES topic's Wikipedia-internal probability and either its Web probability (blue points) or Reddit probability (orange triangles). We observe that certain topics–such as Computing, Medicine & Health, Philosophy & Religion, and Technology–are referenced more frequently on the Web than Wikipedia's internal linking. In contrast, topics like Sports, North America, Northern Europe, Military, and Music appear underrepresented in external links relative to their presence in Wikipedia's internal network. While some topic-level differences in link probabilities may appear numerically small, the scale of our dataset means that even modest differences translate into substantial differences in exposure and referencing across the Web.

Key Findings. Wikipedia article referencing patterns vary across platforms: articles that receive frequent internal links within Wikipedia differ from those that are widely linked on the Web or shared on Reddit. This indicates that social media is not a reliable proxy for broader Wikipedia usage. While previous studies have focused on Reddit and Twitter, our findings suggest that social media linking behavior does not accurately represent how Wikipedia is referenced across the wider Web.

5 Where Articles Are Linked

First, we analyze which types of websites are more likely to link to Wikipedia. To do this, we compare webpages that contain Wikipedia links with a random sample of webpages from the general Web. This allows us to determine whether Wikipedia is disproportionately referenced in certain types of websites.

Second, we examine where within a webpage Wikipedia links appear. We categorize webpage content into three sections–boilerplate, main content, and user contributions–and define HTML-based rules to extract these segments computationally. This helps identify the context and intent behind Wikipedia links.

Finally, we combine these two analyses to explore how different websites use Wikipedia links across different webpage sections. This integrated approach provides a clearer picture of Wikipedia's role in online information dissemination.

Wikipedia vs. the Web

With over 48 million links to English Wikipedia across 940,000 distinct domains, Wikipedia has a strong presence on the Web. Notably, 66% of these domains contain Wikipedia links on multiple webpages, while 3.2% of domains include Wikipedia links on over 100 different webpages. At the individual webpage level, we observe that 3.8 million webpages contain multiple Wikipedia links.

Beyond these aggregate statistics, it is important to characterize the types of websites that link to Wikipedia and

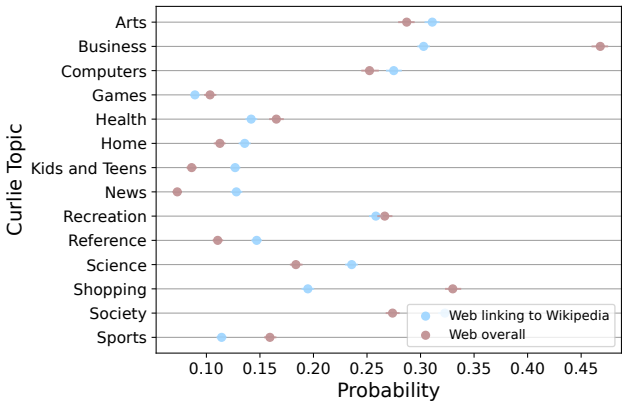

Figure 2: Comparison between websites that link to Wikipedia and a random sample of all websites. For each topic, the difference between the "Web overall" and "Web linking to Wikipedia" probabilities determines its leaning. When the score for "Web linking to Wikipedia" is higher, it implies that the topic tends to link more to Wikipedia, while the reverse implies that Wikipedia articles are underrepresented across that particular topic. Error bars represent 95% bootstrapped confidence intervals.

determine whether certain website categories disproportionately reference Wikipedia. To understand which types of websites link to Wikipedia, we apply simple heuristic rules. Specifically, we define a Wikipedia-related page as any webpage containing "wiki" or "pedia" in its URL, and a blog as any webpage with a date in its URL. Using these rules, we find that 11% of Wikipedia links (4.5 million) originate from encyclopedias or Wiki-style pages, while blogs account for 29% (11.8 million) of all Wikipedia links.

These heuristics provide a basic categorization but do not fully capture the diversity of websites linking to Wikipedia. To improve classification, we use Homepage2Vec, a model that predicts a webpage's topic based on its HTML content. By comparing Homepage2Vec classifications for Wikipedia-linking websites against a random sample of general websites, we quantify which types of websites are most likely to reference Wikipedia.

Implementation and analysis. We apply Homepage2Vec to a sample of 10,000 random domains that link to Wikipedia and 10,000 randomly selected Common Crawl domains. The model assigns a probability distribution across 14 topic categories. Since each webpage can belong to multiple categories, these probabilities do not necessarily sum to 1.

Figure 2 compares the average topic distributions of Wikipedia-linking websites and a random sample of the Web. Our analysis reveals distinct differences between websites that link to Wikipedia and the broader Web. Business and Shopping websites make up a larger portion of the Web but are less likely to link to Wikipedia articles. In contrast, News, Science, Society, Arts, Computers, and Reference websites show a higher tendency to include Wikipedia links.

Segment	Webpages (M)	% of links	Article Entropy	Language Entropy
Boilerplate	3.4	7%	7.6	7.8
Main content	43.7	91%	8.4	8.1
User-Generated	1.0	2%	8.5	7.5

Table 2: Summary of the three segments showing the number of webpages, percentage of all links, article, and language entropy.

Boilerplate	Main content	User-Generated
privacy	share	reply
powered	new	look
twitter	second	actually
yelp	time	thanks
archive	social	think
acknowledge	file	check
copyright	told	link
buy	took	example
exhaustion	available	good
designed	building	interesting

Table 3: The 10 most predictive words characterizing each of the webpage segments.

A qualitative examination of these topics reveals different linking behaviors. News and Science pages often include Wikipedia links within their main content or in user-generated sections–such as on-site comments. In contrast, Business and Shopping websites are more likely to reference Wikipedia in boilerplate content, either linking to their own Wikipedia page or explaining concepts like cookies. These findings motivate the next section of our study, where we analyze how Wikipedia links are embedded within webpage HTML structures.

Webpage Segmentation

Webpages are structured into different segments, each serving a distinct function that influences how links are placed and interpreted (Kohlschütter, Fankhauser, and Nejdl 2010). Segmenting a webpage is critical for understanding the intent behind a link's placement, as different sections are populated with links for specific reasons. For instance, links in boilerplate content typically provide site-wide navigation or general information, whereas links in main content are often used to enrich specific textual explanations. User-generated content, such as comments, includes links added by users in discussions, often for fact-checking or providing additional context.

To systematically analyze link placement, we developed a set of structural rules to identify different webpage segments. Using an iterative approach inspired by grounded theory, we continuously sampled new examples and refined the rules through deliberation among four of the authors, as detailed in Appendix B. Applying these rules to our dataset, we find that Wikipedia links are primarily embedded in main content (91%), while boilerplate content accounts for 7% and user-generated content for 2% (Table 2). This distribution highlights that Wikipedia is most frequently cited as an informational source within core webpage content, rather than in navigation menus or user discussions.

To further understand how different segments reference Wikipedia, we compute the Shannon entropy of Wikipedia articles linked in each segment. Entropy measures diversity, where higher entropy indicates a wider variety of linked articles, while lower entropy suggests a narrower, more predictable set of links. For each segment, we randomly sample 10,000 Wikipedia articles linked from that segment and compute their probability distribution. These entropy values, reported in Table 2, show that boilerplate content has the lowest entropy, suggesting that a small, recurring set of Wikipedia articles is frequently referenced, such as explanations of standard terms like "cookies" or "privacy policy." In contrast, main content and user-generated content have higher entropy, meaning that a more diverse range of Wikipedia articles is linked in these sections.

Since entropy calculations scale with the number of variables, we limit our analysis to 10,000 articles per segment to maintain computational efficiency. Additionally, we focus on relative entropy values rather than absolute ones to effectively compare diversity across segments.

Linguistic comparison. To examine differences in webpage content, we conduct a linguistic analysis of the text surrounding Wikipedia links in each segment. For this analysis, we extract a 150-character context window on both sides of the link while masking the anchor text. We then filter out stopwords, retain the 5,000 most frequent words, and apply a TF-IDF vectorizer to compare word distributions across the three segments.

Table 3 presents the top 10 most predictive words for each segment, identified using a logistic regression model (*one-vs-all*) on TF-IDF representations. The results show notable differences in language use. Boilerplate content is more focused on specific concepts than the other two segments, while user-generated content (e.g., comments) tends to be more conversational, featuring words such as "look", "actually", and "here".

To further analyze variation in textual contexts, we compute the entropy of the language used in each segment. Entropy values, reported in Table 2, reveal that main content has the highest entropy (8.1), followed by boilerplate (7.8) and user-generated content (7.5). This indicates that main content and user-generated content have similar levels of lexical diversity, whereas boilerplate content is more restricted in the variety of words used.

The lower entropy in user-generated content suggests that when Wikipedia links appear in responses or comments, they tend to be referenced in more predictable ways, with limited variation in phrasing. Similarly, the lower entropy in

boilerplate content compared to main content indicates that Wikipedia links in boilerplate sections are used in consistent and repetitive contexts, such as defining terms or linking to standard references.

Topic Variations in Wikipedia Link Placement

The topic of a website is associated not only with its likelihood of linking to Wikipedia but also with where Wikipedia links appear within a webpage. To investigate this, we analyze how different website categories distribute Wikipedia links across boilerplate, main content, and user-generated content (e.g., comments).

This analysis builds on previous observations that boilerplate links are more common in business websites, while user-generated links are more prevalent in forums. To quantify these patterns, we map the Homepage2Vec probability distribution over Curlie topics onto webpage segments, characterizing how different website categories place Wikipedia links. Specifically, for each webpage segment, we compute the mean probability distribution of its predicted Curlie topics (from Homepage2Vec).

Figure 3 illustrates these distributions, revealing notable differences across website categories. The most significant variations occur in Arts, Business, Science, Shopping, and Society websites, suggesting distinct linking behaviors based on the site's purpose.

Among all website categories, Business and Shopping sites are the most likely to place Wikipedia links in boilerplate content. This suggests that while these websites may not reference Wikipedia frequently in their main content, they often use it in boilerplate sections, such as to explain technical terms like "cookies." In contrast, Science and News websites are more likely to place Wikipedia links in main content and user-generated sections, indicating a different usage pattern where Wikipedia serves as a source of additional context or citation within discussions and articles.

These findings highlight that Wikipedia links serve multiple roles across the Web. The key distinction between websites that embed Wikipedia links in boilerplate versus main content or user-generated content appears to be driven by how business-oriented the site is. Business and commercial websites often use Wikipedia for functional or explanatory purposes, whereas informational websites such as Science and News domains rely on Wikipedia more as a content reference.

Key Findings. Wikipedia is disproportionately linked from News, Science, and Society websites, while it is underrepresented in Business and Shopping websites compared to the Web as a whole. When Business and Shopping websites do link to Wikipedia, the references are primarily placed in boilerplate content, often serving functional purposes such as explaining terms like "cookies." In contrast, Science and News websites are more likely to embed Wikipedia links within main content and user-generated content, suggesting that Wikipedia plays multiple roles depending on the website's context. Despite these variations, Wikipedia is most frequently linked from the main content of webpages overall.

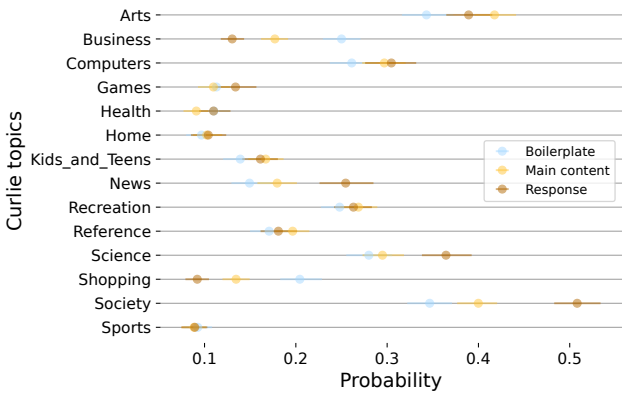

Figure 3: Comparison of where Wikipedia links are likely to occur in a webpage and their Curlie topic. Recall that Homepage2Vec outputs a separate prediction for each Curlie topic, so the probabilities across topics do not necessarily add to 1. Error bars represent 95% bootstrapped confidence intervals.

6 Why Articles Are Linked

Thus far, we have examined what Wikipedia articles are linked and where these links appear. While this provides insights into Wikipedia's presence on the Web, it does not address the intent behind including Wikipedia links. This section explores *why* Wikipedia articles are referenced online.

To determine intent, we conducted iterative coding to classify the motivations behind linking to Wikipedia (cf. Appendix C for details). While there may be a wide range of reasons, our analysis reveals that most links fall into two broad categories: delegation and evidence.

- *Delegation* occurs when a website links to Wikipedia in a manner similar to Wikipedia's internal linking guidelines[6]. These links serve as contextual references, directing users to relevant Wikipedia articles for additional information. Examples include linking to the Wikipedia page for HTTP Cookies when explaining a website's privacy policies or referencing an obscure place or historical figure mentioned in passing. Such links enrich content by providing supplementary explanations, defining technical terms, or offering background information.
- *Evidence* refers to cases where content is sourced directly from Wikipedia. This category includes multimedia elements such as images, audio clips, or videos, as well as text-based references where Wikipedia is cited as the source of a fact, statistic, or quotation. Wikipedia images are commonly linked in biographical or geographic webpages, while text-based references frequently appear in blogs or informational articles where authors cite Wikipedia to support a claim.

These two categories capture the dominant linking behaviors observed across the Web: using Wikipedia as an external reference to delegate explanations or as a source to pro-

[6]https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking#General_points_on_linking_style

vide attribution and supporting evidence. Manually annotating samples, we were able to bucket almost all the Wikipedia sharing links (98%) into these two motivations.

To quantify the prevalence of each type of Wikipedia linking, we randomly sampled 500 webpages that link to Wikipedia and manually annotated each link as either delegation or evidence. Our analysis revealed that delegation is the dominant reason for linking to Wikipedia, accounting for 95% of links, while only 5% are used for evidence. Generalizing these proportions to the 48.8 million Wikipedia links in the main content, we estimate that 46.4 million links serve as delegation, while 2.4 million links are used for evidence.

Key Results. The vast majority of Wikipedia links (95%) are used for delegation, where websites reference Wikipedia to explain a concept. However, evidence citation accounts for 5% of links, representing a small but meaningful portion of Wikipedia's presence on the Web.

7 Discussion

This paper presents the first large-scale dataset of Wikipedia links across the Web, capturing the raw HTML content of webpages that reference Wikipedia articles. The dataset spans a significant portion of the Web, covering links from 1.68% of all domains in the Common Crawl dump. By making this dataset publicly available, we provide a valuable resource for studying how Wikipedia is integrated into the broader information ecosystem.

Summary of Findings

Our analysis reveals several key insights into how Wikipedia is linked across the Web. First, the number of external Web links to Wikipedia articles correlates positively with their in-degree within Wikipedia ($r = 0.5$, $p < 0.001$), indicating that structurally central articles within Wikipedia also tend to attract more external references. At the same time, topic-level distributions of Web links and Reddit citations diverge, suggesting that social media linking behavior is not a reliable proxy for how Wikipedia is referenced across the broader Web.

Second, the types of Wikipedia articles linked on the Web, cited on social media, and structurally significant within Wikipedia differ significantly. The broader Web tends to link to Computing, Medicine & Health, and Technology, while social media platforms such as Reddit are more likely to reference Philosophy and Politics. In contrast, Wikipedia's internal link structure places greater emphasis on Sports, Music, North America, and Northern Europe, reflecting differences between external usage and the internal network structure. As noted in Section 4, differences in units of analysis between the Web and Wikipedia-internal probabilities should be considered when interpreting the magnitude of these effects.

Third, Wikipedia's presence is not uniform across the Web, and different sections of webpages integrate Wikipedia links in distinct ways. Boilerplate content often includes Wikipedia links to define or explain concepts essential to the website, whereas main content links are more common in News websites, where Wikipedia is used to provide contextual references.

Finally, through manual annotation of a random sample, we estimate that Wikipedia links on the Web overwhelmingly serve a delegation function (95%), with a smaller but meaningful share used as evidence (5%). This simple two-category distinction captures the dominant linking behaviors observed across a diverse set of webpages.

Implications and Broader Impact

Our findings reinforce Wikipedia's critical role beyond direct page views, highlighting its function as a form of knowledge infrastructure—serving both as a content source and as a mechanism for delegating explanations across millions of webpages. This underscores the need for the Wikipedia community to maintain and improve the platform, as its influence extends far beyond its own ecosystem. Wikipedia helps shape the dissemination of factual information across the Web, particularly in domains such as news and science, where it is cited most frequently

A central theme of this work is quantifying Wikipedia's value to the broader Web. By better understanding its reach and influence, we can inform strategies for its development to account for diverse user needs and maximize its societal impact.

The release of this dataset opens multiple avenues for future research. It can be used to study the credibility of sources linking to Wikipedia, examine Wikipedia's role in misinformation correction, or track changes in linking behavior over time by integrating additional datasets. Furthermore, it enables investigations into how Wikipedia contributes to knowledge dissemination across different domains and how its presence on the Web evolves in response to external events.

Beyond analyzing Wikipedia's role, this work has major implications for improving Wikipedia itself. Most studies of Wikipedia have focused on isolated analyses of its internal structure, but the WEB2WIKI dataset expands the scope of research by offering a broader, web-scale perspective. Additionally, the methodology presented in this study can be adapted to analyze linking patterns beyond Wikipedia, providing a framework for studying how other entities and websites integrate into the Web's information ecosystem.

Limitations and Future Work

Our study is subject to several limitations. Web data, including Common Crawl, is inherently noisy, making webpage segmentation a challenging task with no universally established approach, even for extracting elements like boilerplate content. We rely on heuristics to segment webpages, which, while not perfect, have been validated through reliability checks to ensure robustness.

Due to practical constraints, we use a single Common Crawl snapshot, meaning our findings may be influenced by data variability across different snapshots. Future work should examine robustness across multiple snapshots to assess consistency over time. Additionally, while Common Crawl is the largest publicly available collection of surface Web data, it excludes platforms that require login access,

such as Twitter, Reddit, or that enforce no-follow policies. As a result, our WEB2WIKI dataset does not capture Wikipedia links from these websites. Also, the datasets compared in Section 4 span different time periods. While Wikipedia's core structure evolves relatively slowly, this temporal mismatch may introduce some noise into topic-level comparisons.

Several promising directions for future research remain. The language surrounding Wikipedia links often reveals unexpected conceptual connections. For instance, links to *Raphael's School of Athens* frequently appear in discussions about the ancient Academy, suggesting lateral relationships between entities (Feith et al. 2024). Similarly, hypertext structures introduce unique human associations, such as the Wikipedia article Proposed referendum on United Kingdom membership of the European Union featuring the phrase "11,757-word behemoth" as hypertext. These contextual relationships could inform link prediction models, RAG training, entity candidate generators (Garcia-Duran, Arora, and West 2022), and entity linking models (Arora, Garcia-Duran, and West 2021; Čuljak et al. 2022), similar to prior work (Spitkovsky and Chang 2012; Singh et al. 2012).

Another promising direction is to investigate the alignment between the number of incoming links, user clickstreams, the importance of these websites, and the actual attention received by linked articles. This would help identify potential mismatches between how often an article is referenced and how often it is accessed, extending prior work on the misalignment between demand and supply of attention on Wikipedia (Warncke-Wang et al. 2015).

Finally, a key direction for future research is to expand the analysis beyond English Wikipedia. While this study focused on English Wikipedia as a starting point, our dataset includes many other language editions. Exploring multilingual Wikipedia linking behavior can reveal cultural differences in citation practices. A preliminary analysis suggests that some languages exhibit far higher view-to-Web-link ratios than others, indicating possible differences in how Wikipedia is perceived across linguistic and cultural contexts. Investigating these patterns could offer valuable insights into global perspectives on Wikipedia's role as a knowledge source.

Ethical Considerations

While the WEB2WIKI dataset is constructed entirely using publicly available resources, the Web is an untamed resource and likely contains personally identifiable information and foul or offensive content. To alleviate this, our data release only consists of the linking structure and the surrounding context as opposed to the entire webpages. Moreover, our dataset does not contain any private or sensitive information about the browsing behavior of readers on the Web. Finally, no author was paid by the Wikimedia Foundation to conduct the analyses presented in this paper, thereby mitigating the potential for conflicts of interest in the analyses.

8 Conclusion

In this work, we introduced the WEB2WIKI dataset and conducted the first large-scale analysis of how Wikipedia is linked across the Web, providing foundational estimates of its presence and influence on the surface Web. Our findings show that 1.68% of all domains (1,475,057) on the surface Web link to Wikipedia, underscoring its global reach and significance.

Wikipedia plays a critical role in the Web's infrastructure, serving as both a de facto external reference for explanations and a frequent source for content and evidence. This highlights its dual function as a knowledge hub and a content provider across diverse online platforms.

The datasets, methods, and insights presented in this study contribute to a deeper understanding of Wikipedia's role on the Web. We hope this work provides a foundation for future research into Wikipedia's impact, as well as broader investigations into web-scale linking structures beyond Wikipedia.

Acknowledgements

We would like to thank Leila Zia, Isaac Johnson, Martin Gerlach, and Manoel Horta Ribeiro for their consultations and valuable inputs. West's lab is partly supported by grants from Swiss National Science Foundation (200021_185043 and 211379), Swiss Data Science Center (P22_08), H2020 (952215), and Microsoft Swiss Joint Research Center. Arora's lab is partly supported by grants from the Novo Nordisk Foundation (NNF24OC0099109), the Pioneer Centre for AI, and EU Horizon 2020 (101168951). We also gratefully acknowledge generous gifts from It-vest - networking universities, Facebook, Google, and Microsoft.

References

Alshomary, M.; Völske, M.; Licht, T.; Wachsmuth, H.; Stein, B.; Hagen, M.; and Potthast, M. 2019. Wikipedia text reuse: Within and without. In *European Conference on Information Retrieval*, 747–754. Springer.

Arora, A. 2024. *Modeling and Enhancing Human Knowledge Navigation*. Ph.D. thesis, EPFL, Lausanne, Switzerland.

Arora, A.; Garcia-Duran, A.; and West, R. 2021. Low-Rank Subspaces for Unsupervised Entity Linking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8037–8054.

Arora, A.; Gerlach, M.; Piccardi, T.; García-Durán, A.; and West, R. 2022. Wikipedia Reader Navigation: When Synthetic Data Is Enough. In *WSDM*, 16–26.

Arora, A.; West, R.; and Gerlach, M. 2024. Orphan Articles: The Dark Matter of Wikipedia. In *Proceedings of the International AAAI Conference on Web and Social Media*, 100–112.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, 830–839.

Čuljak, M.; Spitz, A.; West, R.; and Arora, A. 2022. Strong Heuristics for Named Entity Linking. In *Proceedings of the*

2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, 235–246.

Feith, T.; Arora, A.; Gerlach, M.; Paul, D.; and West, R. 2024. Entity Insertion in Multilingual Linked Corpora: The Case of Wikipedia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 22796–22819.

Forte, A.; Andalibi, N.; Gorichanaz, T.; Kim, M. C.; Park, T.; and Halfaker, A. 2018. Information fortification: An online citation behavior. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, 83–92.

Fortunato, S.; Boguñá, M.; Flammini, A.; and Menczer, F. 2006. Approximating PageRank from in-degree. In *International workshop on algorithms and models for the web-graph*, 59–71. Springer.

Garcia-Duran, A.; Arora, A.; and West, R. 2022. Efficient Entity Candidate Generation for Low-Resource Languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6429–6438.

Halfaker, A.; and 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.

Hinnosaar, M.; Hinnosaar, T.; Kummer, M.; and Slivko, O. 2023. Wikipedia matters. *Journal of Economics & Management Strategy*, 32(3): 657–669.

Holloway, T.; Bozicevic, M.; and Börner, K. 2007. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. *Complexity*, 12(3): 30–40.

Kohlschütter, C.; Fankhauser, P.; and Nejdl, W. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, 441–450.

Lemmerich, F.; Sáez-Trumper, D.; West, R.; and Zia, L. 2019. Why the World Reads Wikipedia: Beyond English Speakers. In *WSDM*, 618–626.

Lugeon, S.; Piccardi, T.; and West, R. 2022. Homepage2Vec: Language-Agnostic Website Embedding and Classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1285–1291.

McMahon, C.; Johnson, I.; and Hecht, B. 2017. The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In *Eleventh international AAAI conference on web and social media*.

Meier, F. 2022. TWikiL–the Twitter Wikipedia Link Dataset. In *ICWSM*, volume 16, 1292–1301.

Mesgari, M.; Okoli, C.; Mehdi, M.; Nielsen, F. Å.; and Lanamäki, A. 2015. "The sum of all human knowledge": A systematic review of scholarly research on the content of W ikipedia. *Journal of the Association for Information Science and Technology*, 66(2): 219–245.

Mitrevski, B.; Piccardi, T.; and West, R. 2020. WikiHist. html: English Wikipedia's full revision history in HTML format. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 878–884.

Moyer, D.; Carson, S.; Dye, T.; Carson, R.; and Goldbaum, D. 2015. Determining the influence of Reddit posts on Wikipedia pageviews. In *Proceedings of the International AAAI Conference on Web and Social Media Workshops*, 75–82.

Piccardi, T. 2022. *How We Use Wikipedia: Studying Readers' Behavior with Navigation Traces*. Ph.D. thesis, EPFL, Lausanne, Switzerland.

Piccardi, T.; Gerlach, M.; Arora, A.; and West, R. 2023. A Large-Scale Characterization of How Readers Browse Wikipedia. *ACM Trans. Web*.

Piccardi, T.; Gerlach, M.; and West, R. 2022. Going Down the Rabbit Hole: Characterizing the Long Tail of Wikipedia Reading Sessions. In *Companion Proceedings of the Web Conference 2022*, WWW '22, 1324–1330. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391306.

Piccardi, T.; Redi, M.; Colavizza, G.; and West, R. 2021. On the Value of Wikipedia as a Gateway to the Web. In *WWW*, 249–260.

Singer, P.; Lemmerich, F.; West, R.; Zia, L.; Wulczyn, E.; Strohmaier, M.; and Leskovec, J. 2017. Why We Read Wikipedia. In *WWW*, 1591–1600.

Singh, S.; Subramanya, A.; Pereira, F.; and McCallum, A. 2012. Wikilinks: A large-scale cross-document coreference corpus labeled via links to Wikipedia. *University of Massachusetts, Amherst, Tech. Rep. UM-CS-2012*, 15.

Spitkovsky, V. I.; and Chang, A. X. 2012. A cross-lingual dictionary for english wikipedia concepts. In *LREC*, 3168–3175.

Taraborelli, D. 2015. The sum of all human knowledge in the age of machines: a new research agenda for Wikimedia. In *ICWSM-15 Workshop on Wikipedia*.

Thalhammer, A.; and Rettinger, A. 2016. PageRank on Wikipedia: towards general importance scores for entities. In *The Semantic Web: ESWC 2016 Satellite Events, Heraklion, Crete, Greece, May 29–June 2, 2016, Revised Selected Papers 13*, 227–240. Springer.

Thompson, N.; Flanagan, B.; Richardson, E.; McKenzie, B.; and Luo, X. 2022. Trial by Internet: A Randomized Field Experiment on Wikipedia's Influence on Judges' Legal Reasoning. In Tobia, K., ed., *Forthcoming in Cambridge Handbook of Experimental Jurisprudence*. Cambridge University Press.

Thompson, N.; and Hanley, D. 2018. Science is shaped by wikipedia: Evidence from a randomized control trial. *MIT Sloan Research Paper*, 5238(17).

Vincent, N.; Johnson, I.; and Hecht, B. 2018. Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Vincent, N.; Johnson, I.; Sheehan, P.; and Hecht, B. 2019. Measuring the importance of user-generated content to search engines. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 505–516.

Vogels, T.; Ganea, O.-E.; and Eickhoff, C. 2018. Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*, 167–179. Springer.

Warncke-Wang, M.; Ranjan, V.; Terveen, L.; and Hecht, B. 2015. Misalignment between supply and demand of quality content in peer production communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, 493–502.

WMF, M. 2022. A New Research Roadmap For Addressing Knowledge Gaps. Available: https://diff.wikimedia.org/2022/04/21/a-new-research-roadmap-for-addressing-knowledge-gaps/, Last accessed: 2023-01-14.

Zia, L.; Johnson, I.; Mansurov, B.; Morgan, J.; Redi, M.; Saez-Trumper, D.; and Taraborelli, D. 2019. Knowledge Gaps–Wikimedia Research 2030.

ICWSM Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. Section 3 introduces the methods, whereas Sections 4, 5, and 6 further provides analysis-specific details.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. Please see Data and Methodology (Section 3).
   (e) Did you describe the limitations of your work? Yes. Please see Section 7.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. Please see Section 7. Overall, we do not foresee any significant negative impact.
   (g) Did you discuss any potential misuse of your work? Yes. Please see Section 7. Overall, we do not foresee any significant negative impact.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. Please see Data and Methodology (Section 3) and Discussion (Section 7).
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? N/A
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? N/A
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N/A
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? N/A
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? N/A
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? N/A
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? No. The collected WEB2WIKI dataset is quite large to be shared with the submission, and sharing via a URL risks breaching the anonymity. Thus, we will release the dataset publicly via Zenodo upon acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? N/A

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. Please see Section 7.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? N/A

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? N/A

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N/A

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? N/A

A Common Crawl Data Collection Policies

To the best of our knowledge, websites with specific "nofollow" or "robots.txt" rules are excluded from the public scrape of Common Crawl. Thus, the WEB2WIKI dataset lacks links from specific proprietary websites like Facebook, Reddit, Twitter, Quora, YouTube, etc., thereby limiting our sample of the Web to the surface web.

B Defining Webpage Segmentation Rules

While there exist many methods such as engineering shallow textual features (Kohlschütter, Fankhauser, and Nejdl 2010) and deep learning techniques (Vogels, Ganea, and Eickhoff 2018) to perform webpage segmentation, here we rely on a simple and intuitive rule-based approach that we propose to segment a webpage into three classes, namely *boilerplate*, *main content*, and *user generated*. This is primarily because of the following three requirements. First, owing to the sheer scale of the WEB2WIKI dataset we need a webpage segmentation model that is efficient and scalable. Second, owing to the lack of readily available ground-truth annotations, the model should be able to learn to segment a webpage using a relatively small set of labeled data samples. Finally, the model should be effective—possessing high precision and recall—and generalizable to unseen data. We note that while the state-of-the-art machine learning approaches mentioned above are effective, they lack efficiency and scalability. On the contrary, the proposed rule-based approach stands strong on all three aforementioned requirements, and is therefore preferred in this work.

To actually define the structural rules, we follow an iterative process inspired by grounded theory, which involves a continuous cycle of sampling Wikipedia links, labeling them, learning the rules by examining the DOM tree, and refining the rules based on intermediate predictions. Four authors were involved in this process. Specifically, we do the following:

1. Sample 1000 random Wikipedia links.
2. Label each link as belonging to one of the three classes, namely boilerplate, main content, or user generated.
3. Learn the rules using the 1000 Wikipedia links (labeled in Step 2) and define a rule-based classification model.
4. Sample 500 random Wikipedia links and apply the learned rules.
5. Refine the rules based on false positives and negatives obtained on the 500 links (sampled in Step 4).
6. Repeat until convergence.

Using the aforementioned iterative process, we learn structural rules that can be grouped into two broad HTML categories, namely tag-based and attribute-based. Then, for each Wikipedia link we iterate through its ancestors in the DOM tree to determine if a specific tag or attribute is included. Let us consider an example of using an attribute-based rule to classify a specific segment of a webpage as boilerplate. We found that "blogroll", which constitutes a set of links the author of a blog post tends to include on the sidebar of their page, is a strong determinant of the boilerplate segment within a webpage. To determine if "blogroll" is an ancestor of an embedded Wikipedia link, we iterate over the ancestors of the Wikipedia link in the DOM tree and conduct a free text search on the class and id attributes of its ancestor nodes for "blogroll". If found, the Wikipedia link is classified as being embedded in the boilerplate segment. A similar process is carried out for all other rules. While clashes in the predictions from rules occur rarely, whenever clashes do occur they are resolved as follows. We first check if the link can be classified to be embedded as a "response". If not, we next check the applicability of the "boilerplate" classification, failing which, the link is classified to be embedded within the "main content" segment of a webpage. Note that the final set of webpage segmentation rules learned using the aforementioned iterative process for the WEB2WIKI dataset is presented in the README of our GitHub repository: https://github.com/blind-anonymous/web2wiki.

We evaluate the efficacy of the learned rules as follows. We sample 100 random Wikipedia links from each class using the predictions from the rule-based model, and label the sampled links to compute precision. Additionally, we sample 500 random Wikipedia links from the WEB2WIKI dataset, apply the rules to obtain predictions, and label the sampled links to measure recall. Table 4 presents the results.

Segment	Precision	Recall
Boilerplate	0.65	0.78
Main content	0.91	0.96
User generated	0.96	0.57

Table 4: Evaluating the efficacy of the proposed rule-based model for determining the segment where a Wikipedia link is embedded within a webpage.

C Defining a Taxonomy for Linking

To define the taxonomy, the authors went through a series of iterative coding exercises to infer intents users may have in including a link to Wikipedia. In general, we found it difficult to infer intent without more direct approaches like surveys. This realization caused us to move away from trying to define a descriptive taxonomy, and instead understand the relationship between Wikipedia and the site that invokes it. Using this approach, we found two fundamental reasons for including a link: (1) delegate an explanation (content enrichment) and (2) attribution either through evidence or content.

We then used Selenium to screenshot a sample of 500 Web links to Wikipedia and then went through a process of manually annotating them. Screenshots were used for annotation since the visual features were a better signal of intent than text alone. In total, 213 articles did not maintain the link or were functioning. For these cases, we relied on the archived HTML in our dataset and rendered the pages from the available code.

D Top domains

Table 5 summarizes the top 30 domains by number of links to Wikipedia. The top linking domains span several distinct categories. Blogging platforms (e.g., wordpress.com, blogspot.com) and hosting services (e.g., appspot.com, free.fr) aggregate links from large numbers of independent content creators. Wiki-adjacent services (e.g., wikiwix.com, wikitrans.net, fandom.com, dbpedia.org) systematically reference or mirror Wikipedia content. News and civic websites (e.g., leparisien.fr, ria.ru, theyworkforyou.com) link to Wikipedia editorially.

At the same time, some high-ranking domains warrant closer inspection: qaz.wiki appears to be a Wikipedia mirror, workers.dev is a Cloudflare serverless hosting endpoint likely reflecting programmatically generated pages, and domains such as coinshome.net, officeequipmenthub.us, and the2010s.com host niche or potentially spam-generated content. These cases illustrate the inherent noise in web-scale crawl data. Importantly, the long-tail structure of the dataset—with over 900,000 domains contributing the majority of links—limits the influence of any single domain on the aggregate patterns reported in the paper.

Rank	Domain	# Links
1	wordpress.com	3149615
2	blogspot.com	2588425
3	free.fr	1205608
4	wikiwix.com	1194518
5	atoogo.com	1146676
6	appspot.com	1051000
7	chinabuddhismencyclopedia.com	1025668
8	wikitrans.net	1016960
9	portalfield.com	914446
10	wikigallery.org	592547
11	fandom.com	557148
12	dbpedia.org	484611
13	iedit.nu	408697
14	coinshome.net	396156
15	pungenerator.org	393664
16	goo.ne.jp	365376
17	qaz.wiki	346522
18	leparisien.fr	316862
19	tsujitadao.jp	312781
20	flintservice.org	295047
21	pskreporter.de	294224
22	theyworkforyou.com	279800
23	ria.ru	276521
24	workers.dev	269546
25	austria-forum.org	269009
26	wikirank.net	258150
27	officeequipmenthub.us	245752
28	syoboi.jp	236344
29	bingj.com	228153
30	the2010s.com	225107

Table 5: Top 30 linked domains by number of incoming links—excluding *wikipedia.org* and *wikidata.org*.