Identifying the Gaps in the Coverage of Web Domains in Wikipedia and Wikidata for Credibility Assessment Purposes

Malinda Lu
malinda.lu@wellesley.edu
Wellesley College

Eni Mustafaraj
eni.mustafaraj@wellesley.edu
Wellesley College

Abstract

In February 2021, Google Search added a new interface feature to support the evaluation of web domains, known as the "About this result" feature. A prominent part of this feature is a snippet of text pulled automatically from Wikipedia, if a Wiki page for the web domain exists. While conducting large-scale audits of Google Search, we discovered that less than 40% of web domains shown in Google Search results contain a Wikipedia page. Then, we retrieved their Wikidata entries and looked at the extent they incorporate features related to W3C credibility signals. The lack of information for many signals points out to avenues for expanding Wikidata coverage.

Keywords: Google Search, Wikipedia, Wikidata, W3C credibility signals, web literacy

Introduction

The importance of Wikipedia in the context of efforts to combat problematic information on the Web has only increased in the past years. When YouTube started to address its conspiracy theories problem in 2018, it chose to do so by introducing information panels underneath its videos, linking to Wikipedia pages.[1] Then, in 2021, Google Search introduced a new feature "About this result", which can be accessed through a discrete vertical ellipsis next to the URL of a search result, as shown in Figure 1. A prominent entry in this feature, displayed in Figure 2, is a text excerpt from the Wikipedia page about the **source** of the information, together with a link to the Wikipedia entry for the source. When promoting this new feature, Google explained that it is intended to help users vet unfamiliar web domains (or sources), following a practice known in the literature as "lateral reading" (Wineburg and McGrew,).

Prior research has established that Google Search results are perceived more valuable when they include Wikipedia entries (McMahon et al., 2017). Furthermore, another feature of Google Search, right-side knowledge panels, which frequently rely on Wikipedia content, are also deemed important when evaluating the credibility of news sources (Lurie and Mustafaraj, 2018). Given this established interdependence between Google Search and Wikipedia, we set out to investigate some aspects of this new literacy-enhancing feature. In prior work, we established that users are still mostly unaware of this feature, because of its minimalist design (Wang et al., 2023). Additionally, they believe that the panel should contain more information about a web domain, especially some of the credibility signals identified by the W3C Community Group on the Credible Web report,[2] such as the organization behind the website, its ownership, awards that it might have received, headquarter's location, etc.

In this abstract, which documents ongoing research, we focus on two aspects of this literacy initiative that are related to the Wikimedia ecosystem: 1) how often does the Google "About this result" link to a corresponding Wikipedia entry, and 2) does its associated Wikidata entry contain structured information that could be mapped to W3C credibility signals?

Methods for Data Collection

To answer our first question on the prevalence of Wikipedia entries in the "About this result" panels, we performed longitudinal audits of Google Search for different topics: political elections, gender expression, and domestic violence. For each topic, we collect search results for several query phrases (ranging from a few dozens to a few hundreds). Our process for auditing Google Search is explained here (Lurie and Mustafaraj, 2019). We use custom code instrumenting the browser automating tool Selenium[3] to automatically click on the vertical ellipsis button next to each search result and extract the content of the pop-up window, including a Wikipedia link, if available. Because this process is time-consuming, we tried to bypass the automated clicking during the audit with a direct Wikipedia search. That is, we searched Wikipedia for the domain names of Google Search results. This is not trivial, since Wikipedia displays multiple results and the first one often relates to sub-organizations or sub-topics instead of the domain's main organization. Therefore, for the second part of the analysis, we only relied on

[1] https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/

[2] https://www.w3.org/community/credibility/

[3] https://www.selenium.dev

the Wikipedia links extracted directly from the Google "About this result" page.

Our second research question investigates the nature of structured information for each web domain in Wikidata. The easiest way we found to access a Wikidata entry was to automatically visit the Wikipedia URL present in Google Search's pop-up panel, and use HTML parsing to extract the Q number from the Wikidata link embedded in the page. Concretely, the Wikipedia page for the EasyChair entry links to `https://www.wikidata.org/wiki/Q1278323` that connects it to its Wikidata record. Once we had the corresponding Q numbers for each domain, we used Wikidata's SPARQL language to collect a complete list of property labels and property values for our web domains.

Results

Performing Google Search audits is not trivial, since Google tries to prevent automated uses of its platform. Having to simulate clicks on each search result's "About this result" button introduces new challenges. For queries related to the 2022 US election, out of the 4,551 unique web domains in the search results, our script clicked on the menus of 2,822 unique domains and found that 1,100 domains (39%) had a Wikipedia link. For the audit that made use of query phrases related to the topics of gender expression and domestic violence (combined together in one audit), we found that 3,856 out of 10,321 unique domains (37%) had a Wikipedia link. A smaller scale audit related to US Supreme Court decisions in June 2022, revealed that out of 363 unique domains appearing in the results, 231 (or 68%) had a Wikipedia link. This high prevalence of Wikipedia links in this latter case points to the nature of the domains, mostly news organizations, which are more likely to have a Wikipedia presence. While this is a good result, the fact that the two other audits, which were larger in scale, displayed a coverage of less than 40%, requires further investigation.

For the 3,856 domains with Wikipedia links found in one of the audits, we analyzed their Wikidata entries (we found 3,758 entries). In Table 1, we summarize the top properties of these entries. Some of them have widespread coverage, providing some useful information, such as the inception date. However, more useful properties, which align more with W3C Credibility signals, such as ownership are less prevalent, as shown in Table 2. Finally, we look at the values of the most common property, `instance of`, shown in Table 2. Some of these values are useful, such as *nonprofit organization* or *open-access publisher*, but others such as *website* or *business* are too generic and need to be accompanied with more specific values to be useful.

Discussion and Future Work

Google continues to rely on Wikipedia for providing information that can help users make sense of what they find in Google Search results. Our study indicates that the coverage of web domains in Wikipedia might be uneven, depending on who is covering a topic. For some topics we studied, electoral information and domestic violence, less than 40% of domains had a Wikipedia page. This might be because these domains lack notability or other issues that need to be uncovered by further research.

Meanwhile, the value of Wikidata information as a source for W3C credibility signals also needs to receive more attention. While there is a large amount of features in Wikidata, they do not map directly to the W3C credibility signals. One important feature, such as the ownership of a website is generally missing, and other signals (e.g. location) are sparsely populated. Futhermore, even for features that are always present, such as `instance of`, the provided values often do not provide discerning information. Collaboration between the W3C Credible Web Community Group and WikiMedia might be necessary to align visions and establish a close collaboration.

Given the importance of information provenance for fighting misinformation on the web, working toward increasing coverage of web domains in Wikipedia and Wikidata is a worthy goal to pursue.

Acknowledgment

The authors acknwoledge the Wellesley Cred Lab members and funding from the NSF IIS 1751087 grant.

References

[Lurie and Mustafaraj2018] Emma Lurie and Eni Mustafaraj. 2018. Investigating the effects of google's search engine result page in evaluating the credibility of online news sources. In *WebSci'18*, pages 107–116.

[Lurie and Mustafaraj2019] Emma Lurie and Eni Mustafaraj. 2019. Opening up the black box: auditing google's top stories algorithm. In *AAAI FLAIRS '19*.

[McMahon et al.2017] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. In *ICWM'17*.

[Wang et al.2023] Ace Wang, Liz Maylin De Jesus Sanchez, Anya Wintner, Yuanxin Zhu, and Eni Mustafaraj. 2023. Assessing google search's new features in supporting credibility judgments of unknown websites. CHIIR '23, page 303–307.

[Wineburg and McGrew] Sam Wineburg and Sarah McGrew. Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teachers College Record*.

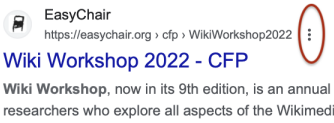

Figure 1: Each URL of a Google search result is accompanied by a vertical ellipsis, which when clicked opens a pop-up window for a feature known as "About this result". See Figure 2 for an example of this feature.

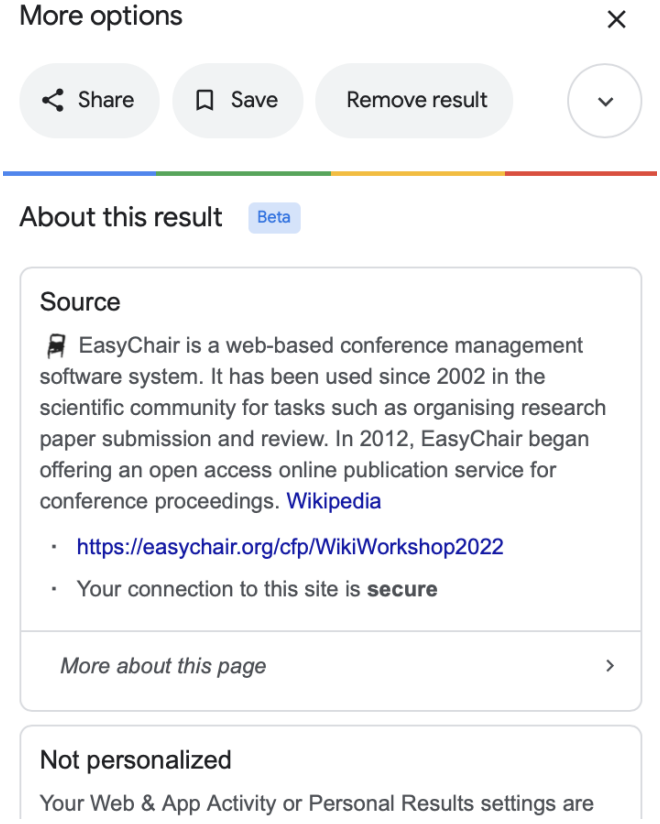

Figure 2: The "About this result" feature for "wiki workshop". The first box, labeled "Source" displays a text snippet that is lifted from the Wikipdia page of the domain `easychair.org`.

Table 1: Top properties of Wikidata entries ranked by occurrence.

Property	# of Occur.	% of Entries
instance of	3724	99.1
official website	3601	95.8
Freebase ID	3400	90.5
inception	3308	88.0
Twitter username	2936	78.1
country	2893	77.0
social media followers	2746	73.1
headquarters location	2019	53.7
Instagram username	1828	48.6
Library of Congress ID	1715	45.6

Table 2: Other useful properties in Wikidata entries, relevant to W3C Credibility signals.

Property	# of Occur.	% of entries
founded by	993	26.4
owned by	810	21.6
award received	232	6.2
publisher	201	5.3

Table 3: Top values for the 'instance of' property in our dataset of web domains.

Property Value	# of Occur.	% of prop.
website	633	17.0
business	516	13.9
public educational institution in the US	329	8.8
university	276	7.4
organization	220	5.9
social networking service	217	5.8
nonprofit organization	200	5.4
open-access publisher	191	5.1
magazine	190	5.1