Insights on the references of Wikipedia's featured articles in English, French, Portuguese and Spanish

Nidia A. Hernández
CAICYT-CONICET
Argentina
nidiahernandez@conicet.gov.ar

Gimena del Rio
IIBICRIT/HD CAICYT-CONICET
Argentina
gdelrio@conicet.gov.ar

Diego de la Hera
Independent Researcher
Argentina
delahera@gmail.com

1 INTRODUCTION

References play an essential role for Wikipedia's reliability. Statements must be supported by external sources allowing the reader to verify the validity of an article. However, identifying and ranking the references used in Wikipedia is not a trivial task, therefore efforts for analyzing and constituting datasets of citations are of considerable value [2, 4]. On the one hand, references can't be requested through the Mediawiki's API, so the only way to retrieve them is extracting them from the wikitext of each article. On the other hand, references can be entered either by hand or using a template since the citing guidelines leave free choice: "the use of citation templates is neither encouraged nor discouraged"[1]. In fact, a recent study [1] about reference editing practices among Wikipedia's experienced editors confirms the diversity of practices when editing references: 32% of editors do not use tools when creating articles and 5% use only offline resources for references.

Wikipedia's Visual Editor allows to automatically generate citations from URLs via an extension called Citoid. The citations produced by Citoid for sites which embed the metadata appropriately are correct. This is often the case for libraries, scientific journals, repositories. However, the results are often inaccurate or incomplete for non academic sites such as newspapers [3]. We will refer to this as the Citoid's coverage gap.

Web2cit[2], a tool aiming to improve the results for web sources, is currently under development. In this abstract, we report on our first steps for automatically estimate the width and nature of this gap, both before and after the implementation of this tool.

In the following sections we will describe the methodology for the evaluation of the accuracy of automatically generated references for web sources, we will explain the first findings and share some insights regarding cultural differences in editorial practices in English, French, Portuguese, and Spanish Wikipedia.

2 METHODOLOGY

To estimate Citoid's coverage gap for different language Wikipedias, accurate citation metadata is required to compare against Citoid's results. For this reason, we decided to work with featured articles. In order to access this category, articles are submitted to an editorial review process which evaluates, among other criteria, the appropriateness and the formatting of the citation[3]. Hence, we assume that their citation metadata is curated and overall correct.

Our corpus consists of 10.5k featured articles retrieved with Wikipedia's action API[4]. The selection covers Wikipedias in four different languages: English (~6k featured articles), French (~2k featured articles), Portuguese (~1.3k featured articles) and Spanish (~1.2k featured articles).

The wikicode of these featured articles is parsed to identify the citation templates used on each article. Citation templates contain a set of parameters allowing to describe the source: author(s), title, URL, publisher, website, journal, etc. The name of these parameters may vary according to the type of source (web, book, news, journal, thesis, etc.) and the language. We are only interested on the citation templates with a URL.

The citation templates available depend on the language: English Wikipedia has a vast list including over 100 source-specific templates while the generic templates for Spanish Wikipedia barely surpass 50. The names and the parameters of the templates also vary from one language to another, for instance: cite news (en), cita noticia (es), article (fr), citar jornal (pt); publisher (en), periódico (es), périodique (fr), jornal (pt). For this reason, our identification of citation templates is based on a manually curated list[5]. In this work we focused on a subset of parameters that map to a limited set of metadata. The names and mapping of these parameters is also based on our curated list.

Additionally, we quantify all the references counting the <ref> tags. These references may be created with or without a citation template. Repeated references are considered as follows: if a source is cited several times in one article, it is counted only once, but if the same source is cited in, for instance, 3 articles, it is counted 3 times. This is because reference metadata may differ across articles.

[1] https://en.wikipedia.org/wiki/Wikipedia:Citing_sources#Citation_templates_and_tools
[2] https://meta.wikimedia.org/wiki/Grants:Project/Diegodlh/Web2Cit:_Visual_Editor_for_Citoid_Web_Translators
[3] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles
[4] https://{lang}.wikipedia.org/w/api.php?
[5] https://docs.google.com/spreadsheets/d/1xbc3FKE0m4JQHa6WCXtBbzeJ9in8P0EQ2NF_VNsaBaM/edit#gid=0

3 FINDINGS

From the total of 10.5k featured articles, we found 968k references and 461k citation templates with a URL and mentioned on our curated list of template names (see **Figure 1**). On this point, we must highlight that many articles contain sections such as "Further reading" or "Bibliography" where citation templates are not surrounded by <ref>. This is rendered in the web browser as bibliographical references that are not numbered and do not point to a specific text in the article's body.

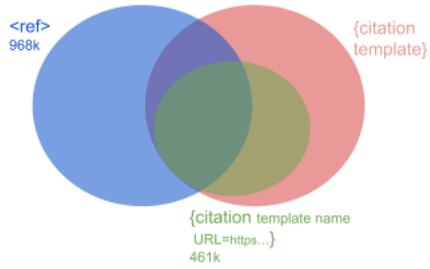

Figure 1: References and citation template shift. The blue circle represents the citations enclosed by <ref> tags, the pink circle represents the citations entered with a citation template, and the green circle, the subset of citation templates with a URL mentioned in our list. Area sizes are indicative and do not represent the actual numbers

The sample shows an average of 92 references and 47 citations templates per article, this could collaborate with the assumption that the featured articles represent highly reliable articles. The use of citation templates is widely generalized, since 94% of the sample present at least one citation template.

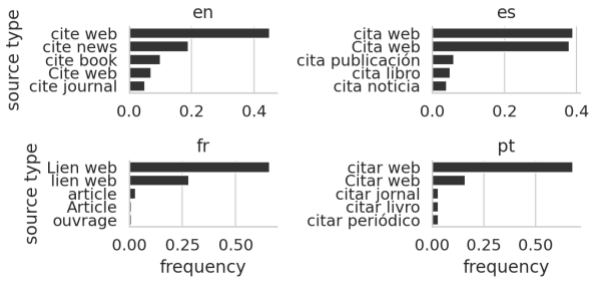

Figure 2: Citation template frequency

Regarding citation templates in particular, **Figure 2** summarizes the names of the top 5 citation templates that were identified, and how many instances of each were extracted. The templates for web sources are the overwhelming majority, representing almost 80% of the citation templates in Spanish, French and Portuguese. The template for newspapers sources (cite news) shows a relevant value only for English.

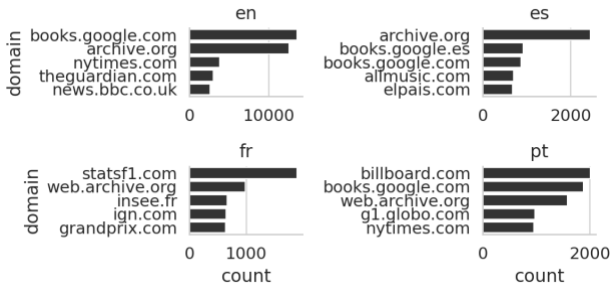

Figure 3: Top domains by language

The main domain names for each language allow to infer cultural differences in editorial practices, as can be seen in **Figure 3**. English only presents domains related to digitized books (books.google.com, archive.org) and newspapers, meaning that the articles refer to general interest and current affairs, while other Wikipedias show a different pattern, with more specific domains for the citations, such as car racing stats (statsf1.com, grandprix.com) and video games (ign.com) in French Wikipedia, and music sources (billboard.com, allmusic.com) in the top places for Portuguese and Spanish. News sites (globo.com, elpais.com, from O Globo and El País, major newspapers in Brazil and Spain respectively) also rank high for these languages in spite of the low values for cita noticia and cita periódico. We presume that the template for web sources is used for online newspapers. Finally, the presence of archiving platforms (web.archive.org) in Portuguese and French Wikipedias suggests that archiving websites is a widespread practice in those communities.

The next step in our research will be comparing these presumably accurate metadata extracted from featured articles' references against Citoid results for the corresponding sources, to estimate the current Citoid's coverage gap and evaluate how this may change with the upcoming deployment of Web2Cit.

REFERENCES

[1] Lucie-Aimée Kaffee and Hady Elsahar. 2021. References in Wikipedia: The Editors' Perspective. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia*. ACM, New York, NY, USA, 535-538. DOI: https://doi.org/10.1145/3442442.3452337

[2] Włodzimierz Lewoniewski, Krzysztof Węcel, and Witold Abramowicz. 2017. Analysis of References Across Wikipedia Languages. In Information and Software Technologies, Robertas Damaševičius and Vilma Mikašytė (eds.). Springer International Publishing, Cham, 561–573. DOI: https://doi.org/10.1007/978-3-319-67642-5_47

[3] Andrew Lih and Rob Fernandez. 2017. Citoid performance for news citations. In WikiCite 2017. Vienna.

[4] Harshdeep Singh, Robert West, and Giovanni Colavizza. 2021. Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies* 2, 1 (April 2021), 1–19. DOI: https://doi.org/10.1162/qss_a_00105