Full paper

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

OPEN ACCESS

Citation: Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Where is the science in Wikipedia? Identification and characterization of scientifically supported contents. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22103). https://doi.org/10.5281/zenodo.6967465

Collection: https://zenodo.org/communities/sti2022grx/

Where is the science in Wikipedia? Identification and characterization of scientifically supported contents

Wenceslao Arroyo-Machado[*], Daniel Torres-Salinas[*] and Rodrigo Costas[**]

[*]*wences@ugr.es; torressalinas@ugr.es*
Department of Information and Communication Sciences, University of Granada, Spain

[**]*rcostas@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands
DST-NRF SciSTIP, Stellenbosch University, South Africa

Introduction

One of the most successful and well-known products of the Web 2.0 is Wikipedia. From its beginnings in 2001 until now it has maintained its essence, a free online encyclopedia that anyone can edit, expanding its content around the world, currently reaching 315 active language editions[1]. Wikipedia has emerged as a global information source from a social and collaborative construction of knowledge. Its open nature has sometimes been accompanied by issues and challenges that have endangered its contents and generated distrust, like the lack of accuracy and readability, vandalism or the presence of bots. However, all these problems have been timely addressed and, over the years, the popularity of Wikipedia has improved. The use of Wikipedia in educational environments is increasingly promoted (Soler-Adillon, Pavlovic, & Freixa, 2018), vandalism prevention systems have been proposed (Martinez-Rico, Martinez-Romo, & Araujo, 2019), and bots have proven to be very useful in tasks such as fixing links and contents, providing suggestions or detecting policy violations (Zheng, Albano, Vora, Mai, & Nickerson, 2019).

The quantitative study of Wikipedia

Beyond the analysis of its limitations and challenges, Wikipedia has been also widely studied from many different fields of knowledge, including several scientometric studies. For example, Colavizza (2020) determined the coverage of COVID-19 literature on Wikipedia; Torres-Salinas, Romero-Frías, and Arroyo-Machado, (2019) adapted traditional bibliometric science maps to Wikipedia; and Jemielniak, Masukume, and Wilamowski (2019) made a ranking of medical journals based on Wikipedia citations. Arroyo-Machado, Torres-Salinas, Herrera-Viedma, and Romero-Frías (2020) studied the subjects of the scientific publications cited in Wikipedia, identifying an overrepresentation of citations to biomedical publications. Most of these previous studies took a rather classical "altmetric" perspective, in which the main object of analysis were the scientific publications cited on Wikipedia.

[1] https://en.wikipedia.org/wiki/List_of_Wikipedias (Accessed on 29 March 2022)

However, following the "heterogeneous couplings" theoretical framework (Costas, Rijcke, & Marres, 2021), it is also possible and relevant to study the specific interactions that Wikipedia has with science (e.g. how science is being used and portrayed on Wikipedia). These types of analyses are much scarce, and there have been fewer scientometric studies that took Wikipedia pages as their main object of analysis and studied their use of scientific publications. One of these few studies was by Yang and Colavizza (2022), who identified the main areas of knowledge of Wikipedia pages citing scientific publications.

Classification challenges in Wikipedia

Unlike other encyclopedias, Wikipedia does not count with any pre-established taxonomy under which Wikipedia pages can be thematically structured. The existing categorization system in Wikipedia is a folksonomy[2] with a complex hierarchical system and with a tangled structure of relationships (Minguillón, Lerga, Aibar, Lladós-Masllorens, & Meseguer-Artola, 2017). Previous proposals to classify Wikipedia (Bairi, Carman, & Ramakrishnan, 2015) or identify topics, such as the Wikidata taxonomies (Mittermeier, Correia, Grenyer, Toivonen, & Roll, 2021) or ORES' page topic model (Yang & Colavizza, 2022), didn't provide comprehensive solutions to the classification and identification of main thematic areas on Wikipedia. The lack of such classification system hinders the possibility of studying Wikipedia from a general thematic perspective, and particularly the study of the presence of scientific literature across Wikipedia thematic areas.

Aim of the study

The main ambition of this study is to explore and identify what Wikipedia thematic areas exhibit a higher use of scientific publications in their references. To achieve this ambition, it is imperative to map a comprehensive thematic landscape of Wikipedia. To approach such mapping challenge, we aim first at exploring the possibilities of using Wikipedia categories and WikiProjects as potential sources of information to categorize and map the main thematic areas on Wikipedia. Once the thematic landscape of Wikipedia is discussed, we aim at identifying those Wikipedia thematic areas with more references to scientific publications and books.

Methodology

Data

We used the Wikipedia Knowledge Graph dataset (Arroyo-Machado, Torres-Salinas, & Costas, 2022). This is a complete and curated dataset of the English edition of Wikipedia, in which different data dumps offered by Wikipedia itself and other datasets elaborated and openly shared by other researchers in open repositories are combined. It includes data on Wikipedia pages, their categories, links between pages and with categories, as well as the bibliographic references of the pages. Note that there are 53,710,529 pages[3], but only 6,328,134 are encyclopedic articles (hereinafter referred to as "Wikipedia pages"). All data included are as of July 2021, except for the bibliographic references, which are as of May 2020 (Singh, West, & Colavizza, 2021).

For the thematic classification of the Wikipedia pages, we first explored Wikipedia categories. Wikipedia distinguishes between content and administrative categories. Both are clearly distinguishable as the former are present in the footer of Wikipedia pages, while the latter are hidden. Wikipedia has a total of 2,064,261 categories, of which 30,146 (1.44%) are hidden, but

[2] Wikipedia editors, known as wikipedians, freely assign one or more "categories" to individual pages.

[3] Other types of pages that are not considered "encyclopedic articles" are user, file, template, help, category or talk pages.

not all of them are used[4]. When considering the categories present in at least one Wikipedia page this number is reduced to 1,473,628 thematic categories tagged in 6,327,692 Wikipedia pages and establishing 35,608,917 links between categories and pages.

As an alternative to Wikipedia categories, we also explored WikiProjects[5] as a source of labels to thematically study Wikipedia pages. WikiProjects is a community of wikipedians that work together to improve the Wikipedia contents around one specific "theme", such as the WikiProjects "Culture"[6], "COVID-19"[7] or "The Beatles"[8] (see Appendix 1 for a more detailed description of WikiProjects). Since the project names are short and precise, there are a small number of projects (a total of 2,118), they contain a curated list of theme-related pages, and 5,522,676 pages of all Wikipedia pages (87.27% of the total) are in at least one project. Thus, WikiProject names can be used also as Wikipedia page descriptors. WikiProjects have a hierarchical structure, much simpler than that of Wikipedia categories. Given this simplicity, the WikiProjects names at the highest levels of the hierarchy are used in the ORES predictive model[9], thus reducing the classification to a few thematic areas or a combination of them. In our case, different from the ORES model we use all the project names as thematic labels.

We retrieve the WikiProject name using the categories, also identifying the quality assigned to that page within the project, so we intend to classify the most relevant Wikipedia pages as representative of their content rather than all of them. We only consider the highest quality classes (Featured, A-class, Good, B-Class, and C-class) because their articles not only may be scarce in content but also offer few references according to their descriptions[10]. The list classes (Featured List and List) were also omitted because lists are page indexes and not articles. Thus we have a sample of 525,390 pages (8.30% of the total number of pages) that are in a WikiProject with an associated quality level that ranges between Featured and C-class.

Finally, in order to identify the scientific literature referenced in Wikipedia pages, the number of references that included a DOI or an ISBN were calculated separately (all other references including media or other types of outputs were not considered in this study). In this study and for the shake of simplicity, DOIs were used as a crude proxy to identify scholarly papers, while ISBN were merely separated as a way to identify books. In those cases where books had both DOI and ISBN they were only counted in the ISBN group (i.e. as books).

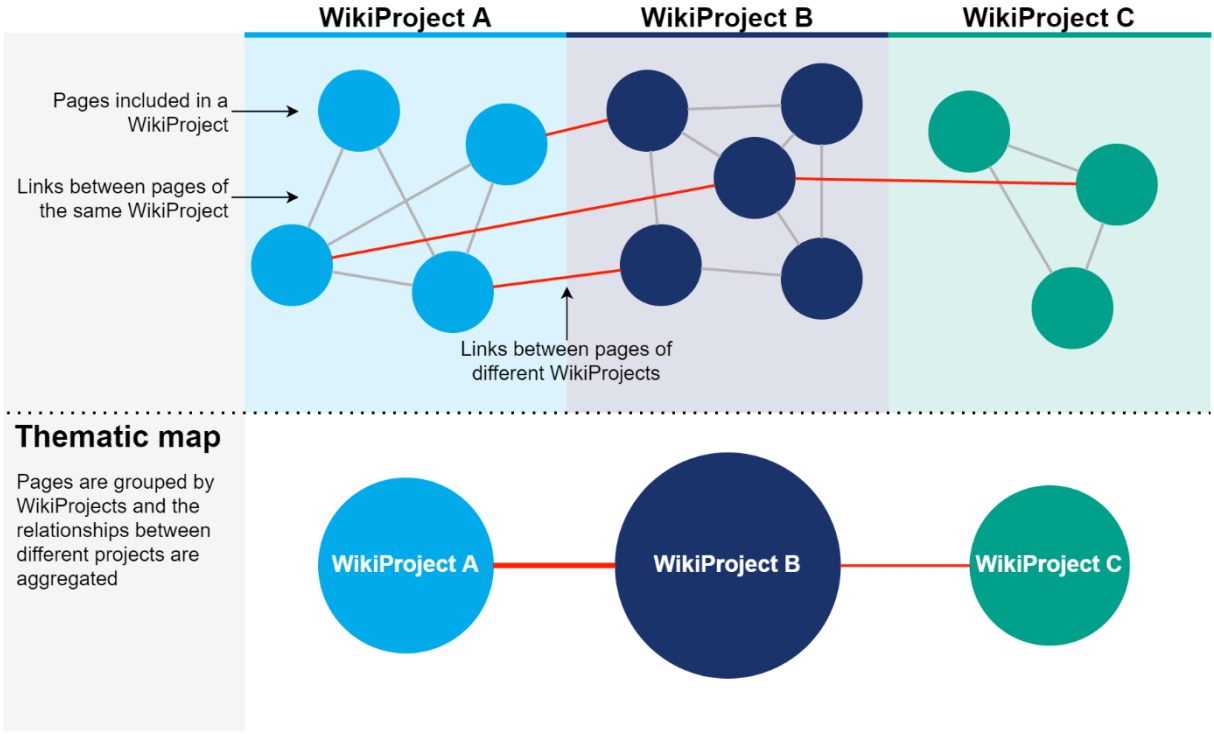
Methods

In order to identify the main thematic areas of Wikipedia, two different approaches have been addressed. Firstly, the distribution Wikipedia categories used in Wikipedia pages was studied, identifying the most common used Wikipedia categories. Secondly, a thematic map was created using WikiProject names. This process is summarized in Figure 1. The WikiProject names of the Wikipedia pages with the highest quality levels were selected as thematic areas, and the WikiProjects were connected based on the aggregation of the links among the Wikipedia pages. Although these links among Wikipedia pages reflect directional relationships (e.g. Wikipedia page X links to Wikipedia page Y, but not vice versa), in this study the links were simplified

[4] There are a total of 206,085 categories that are not assigned to any page, mainly because other more specific categories are created and redirected to them, while other categories are used to group pages that are not encyclopedic articles, such as talk pages.
[5] https://en.wikipedia.org/wiki/WikiProject
[6] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Culture
[7] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_COVID-19
[8] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_The_Beatles
[9] To predict the list of hierarchical higher levels topics that correspond to a given Wikipedia page.
[10] https://en.wikipedia.org/wiki/Wikipedia:Content_assessment (Accessed on 29 March 2022)

so that two WikiProjects are connected by undirected links regardless of directionality. The representation of this network of WikiProject names was carried out using VOSviewer, identifying through its clustering algorithm WikiProjects communities and through overlays those with the highest presence of referenced publications based on the average number of DOIs and ISBNs referenced in their pages.

Figure 1. Methodological proposal for the classification of Wikipedia pages based on WikiProjects and links between pages



The data processing has been carried out in Python and R, all the scripts are available as Jupyter Notebooks on GitHub: 10.5281/zenodo.6445676.

Results

Wikipedia thematic areas

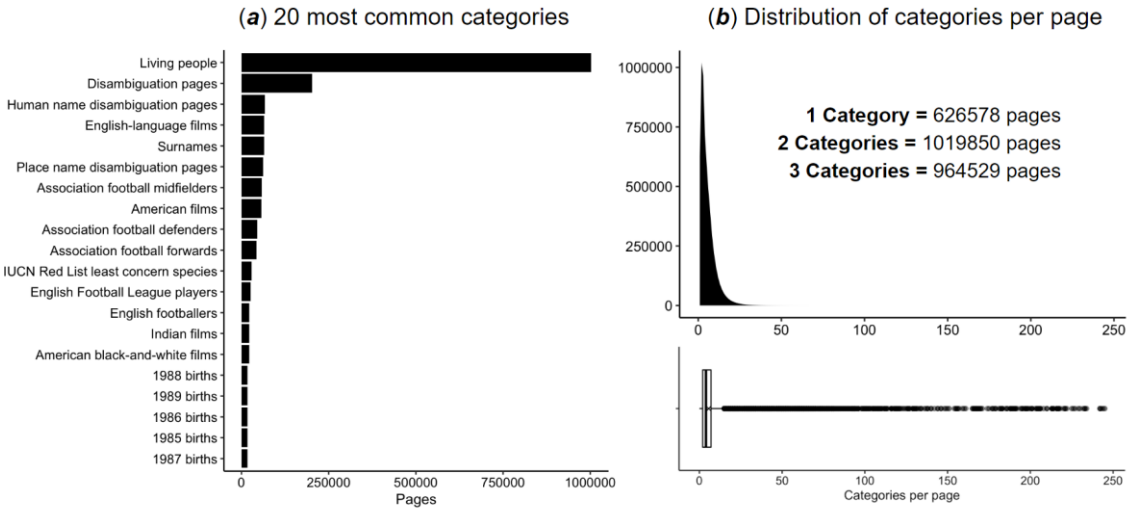
Figure 2a illustrates the main distribution of Wikipedia categories across Wikipedia pages. As it can be seen, they show a very skewed distribution, with some categories being massively used by many Wikipedia pages, while other are hardly use in any. The most used Wikipedia category is "Living people", used by 1,002,407 (15.84%) pages, while in second and third place we find the categories "Disambiguation pages" and "Human name disambiguation pages", with 201,989 and 67,378 pages respectively[11]. Clearly, the thematic descriptive capacity of these three top categories is very limited[12]. Figure 2b illustrates a different perspective of the problem in the skewness of the use of Wikipedia categories. Overall, Wikipedia pages included a mean of 5.59 categories (±4.77) per page, but most Wikipedia pages used just one (1,019,850 pages, 16.12%) or two (2,610,957 pages, 41.26%) categories. This becomes a problem when

[11] Even though these two categories do not have an administrative approach, like the case of the hidden categories, they cannot be related to actual thematic areas, being more descriptors of the type or functionality of the Wikipedia pages.

[12] The range in the specificity of Wikipedia categories ranges between very specific as "Bibliometricians" (11 pages) to very general ones as "Surnames" (64,662 pages).

considering the large number of categories that exist and their granularity[13]. These exploratory analyses suggest that Wikipedia categories are a limited tool to thematically classify Wikipedia.

Figure 2. 20 most common Wikipedia categories and distribution of categories per Wikipedia page.



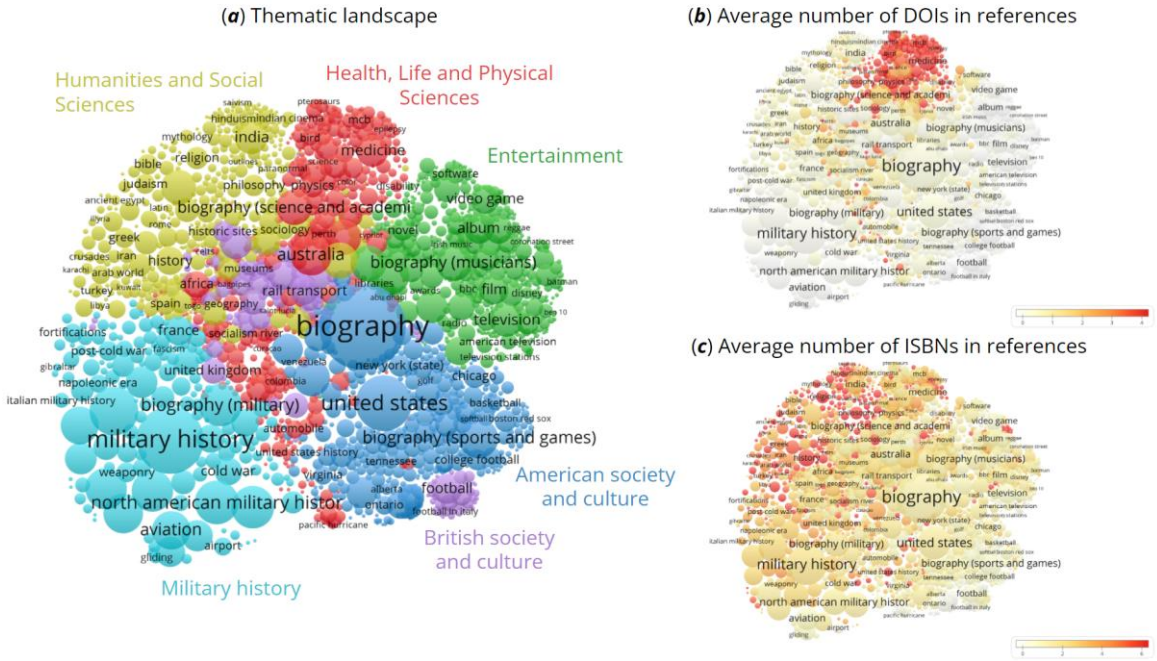
WikiProjects

As mentioned above, in this paper we argue that WikiProjects, and particularly WikiProject names can be used as thematic labels. This is illustrated in Figure 3a, where a general Wikipedia thematic landscape is extracted from with WikiProject names. The total of 2092 WikiProjects (only 26 projects do not have any pages at the highest quality levels) are organized into six clusters based on the direct linkages among Wikipedia pages across WikiProjects. The six main clusters identified broadly correspond with (clockwise) "Humanities & Social Sciences" (yellow), "Health, Life and Physical Sciences" (red), "Entertainment" (green), "American society & culture" (blue), "British society & culture" (purple") and "Military history" (cyan).

There are a few remarkable aspects observable in the map. For example, on the other border of Natural Sciences (red) and Entertainment (green) there are many Wikipedia pages about technological thematic areas such as the Software project. Moreover, the WikiProjet on biographies (node “biography” in blue) occupies a quite important role in the overall Wikipedia, being transversal to many of the other thematic areas, which resonates with the fact that the most used Wikipedia category is “Living people” as discussed above. Another remarkable aspect is the abundance of content related to military history, which concentrates a large number of pages. A similar observation can be made with regards to the clusters of WikiProjects on American and British societies and culture, which may indicate that the choice of the Wikipedia in English language may introduce potential biases towards thematic areas relevant to anglophone communities.

[13] This can be illustrated with the "Bibliometrics" category, as it only includes 15 Wikipedia pages, while the Wikipedia pages of "Samuel C. Bradford", "Eugene Garfield" or "Derek J. de Solla Price" instead of being tagged with the category “Bibliometrics”, they are tagged in the "Bibiometricians" category. The hierarchy problems discussed makes it difficult to work with them.

Figure 3. (a) Thematic landscape of English edition of Wikipedia based on WikiProjects and average number of referenced (b) DOIs and (c) ISBNs



An interactive version can be accessed at:
https://app.vosviewer.com/?map=https://raw.githubusercontent.com/Wences91/wikipedia_science/main/results/wikiprojects_map_vos.txt&scale=1.33&item_size_variation=0.38&score_colors=White-yellow-orange-red

Finally, although clearly the use of WikiProject represents an interesting approach to study the overall Wikipedia thematic landscape, the use of WikiProjects is far from perfect and some problems and challenges remain[14], so this work will be continued to improve the classification and also to include more Wikipedia pages.

Use of DOIs and ISBNs by Wikipedia thematic areas

Figures 3b) and 3c) illustrate the presence and use of DOIs and books across the Wikipedia thematic landscape. Figure 3b) shows how the Wikipedia pages from the "Health, Life and Physical sciences" cluster exhibit the highest average number of DOIs referenced. As with books, Wikipedia thematic areas with a clear humanistic and social sciences component (roughly the "Humanities and Social sciences" cluster) show the largest numbers of references to books with ISBN. This distinction in the use of published material by the two more academic clusters of Wikipedia may be a reflection of the differences that exist in their patterns of scientific communication and consumption of scientific literature also in the social sciences and humanities and in the natural sciences (Hicks, 2005).

Concluding remarks

There are several conclusions that can be extracted from this study that can pave the way towards a future scientometric research program on Wikipedia studies.

[14] For example, the projects "Medicine/Translation task force" and "Months in the 1900s" have the highest average numbers of DOIs (49.52) and ISBNs (33.00) referenced, but the former is focused on translating medical content and the latter on any page associated with that temporal period. Similarly, some projects are no longer active or are undergoing changes, such as the "Medicine/Translation task force", which ceased to be a project to create its own wiki.

From a conceptual point of view, this study demonstrates how Wikipedia can be studied from a quantitative perspective, but taking Wikipedia as the main object of analysis, enabling the study of the role of science on Wikipedia (and its thematic areas) instead of the mere study of the impact of those publications mentioned on Wikipedia (as in the more classical "altmetric" approaches).

This study illustrates the challenges of developing a broad Wikipedia thematic landscape. Particularly the limitations of Wikipedia categories in providing an overview of the thematic areas covered in Wikipedia are shown. The use of WikiProjects is presented as a viable although limited alternative, providing interesting classificatory possibilities. The classification proposed here can be useful for further research on Wikipedia as well as for other researchers who want to identify Wikipedia dynamics in a more aggregated and visual way[15].

The overall thematic landscape of the English language Wikipedia pages developed in this study shows how geographical and cultural singularities emerge (e.g., with a higher presence of Anglo-Saxon thematic areas), as well as thematic areas biases (e.g. the strong presence of military history Wikipedia pages). The identification of such thematic singularities and biases across different Wikipedia editions is relevant not only for Wikipedia itself, but also for altmetric researchers, since social media metrics based on Wikipedia citations may also be influenced by such singularities and biases.

Finally, the analysis of the use of scholarly material on Wikipedia also contributes to the better understanding of the role of science in the development of encyclopedic knowledge, as well as to the study of how epistemic practices (e.g. the differentiated use of scholarly material in the Social Sciences & Humanities and in the Natural Sciences) also arise in their encyclopedic representation.

References

Arroyo-Machado, W., Torres-Salinas, D., & Costas, R. (2022). Wikipedia Knowledge Graph dataset. Zenodo. Retrieved March 23, 2022, from https://doi.org/10.5281/zenodo.6346900

Arroyo-Machado, W., Torres-Salinas, D., Herrera-Viedma, E., & Romero-Frías, E. (2020). Science through Wikipedia: A novel representation of open knowledge through co-citation networks. *PLOS ONE*, *15*(2), e0228713.

Bairi, R., Carman, M., & Ramakrishnan, G. (2015). On the Evolution of Wikipedia: Dynamics of Categories and Articles. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(5), 6–10.

Colavizza, G. (2020). COVID-19 research in Wikipedia. *Quantitative Science Studies*, 1–32. MIT Press.

Costas, R., Rijcke, S., & Marres, N. (2021). "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, *72*(5), 595–610.

[15] For example, the overlay maps of the overall Wikipedia thematic landscape could be explored using other indicators such as the number of visits, editions or discussions, expanding the applications of this study.

Hicks, D. (2005). The Four Literatures of Social Science. En H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 473-496). Dordrecht: Kluwer Academic Publishers.

Martinez-Rico, J. R., Martinez-Romo, J., & Araujo, L. (2019). Can deep learning techniques improve classification performance of vandalism detection in Wikipedia? *Engineering Applications of Artificial Intelligence*, *78*, 248–259.
Minguillón, J., Lerga, M., Aibar, E., Lladós-Masllorens, J., & Meseguer-Artola, A. (2017). Semi-automatic generation of a corpus of Wikipedia articles on science and technology. *Profesional de la Información*, *26*(5), 995–1005.

Mittermeier, J. C., Correia, R., Grenyer, R., Toivonen, T., & Roll, U. (2021). Using Wikipedia to measure public interest in biodiversity and conservation. *Conservation Biology*, *35*(2), 412–423. John Wiley & Sons, Ltd.

Singh, H., West, R., & Colavizza, G. (2021). Wikipedia citations: A comprehensive data set of citations with identifiers extracted from English Wikipedia. *Quantitative Science Studies*, *2*(1), 1–19.

Soler-Adillon, J., Pavlovic, D., & Freixa, P. (2018). Wikipedia in higher education: Changes in perceived value through content contribution. *Comunicar*, *26*(54), 39–48.

Torres-Salinas, D., Romero-Frías, E., & Arroyo-Machado, W. (2019). Mapping the backbone of the Humanities through the eyes of Wikipedia. *Journal of Informetrics*, *13*(3), 793–803. Retrieved April 11, 2022, from https://linkinghub.elsevier.com/retrieve/pii/S1751157718302955

Yang, P., & Colavizza, G. (2022). A Map of Science in Wikipedia. *ArXiv:2110.13790 [cs]*. Retrieved April 3, 2022, from http://arxiv.org/abs/2110.13790

Zheng, L. (Nico), Albano, C. M., Vora, N. M., Mai, F., & Nickerson, J. V. (2019). The Roles Bots Play in Wikipedia. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3359317

Appendix 1. Thematic area identification via WikiProjects

WikiProjects communities do not only identify the pages related to one thematic area but they also catalog them by levels of quality (Featured, A-class, Good, B-Class, C-class, Start-class, and Stub-class) and importance (Top, High, Mid, and Low). A page can be in more than one WikiProject and with different levels of quality and importance in each one. Any wikipedian can carry out these assignments, that are normally reflected by categories on the talk page, where wikipedians discuss the contents of the article. This category can be present mainly in two ways: the term WikiProject with the name of the project or the level of quality or importance with which that page has been assigned in the project followed by its name. For example, in the talk page of "Derek J. de Solla Price" we find both the category "WikiProject History of Science articles" that indicates its belonging to that project and "Start-Class history of science articles" that reflects that for that project it has been marked as Start-Class quality level.