Journal of open research software

WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction

SOFTWARE METAPAPER

PASCHALIS AGAPITOS
JUAN-LUIS SUÁREZ
GUSTAVO ARIEL SCHWARTZ
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

WikiTextGraph is an open-source Python package designed to extract and process text from Wikipedia dumps and construct internal link networks across multiple language editions. It uses efficient parsing, redirect resolution, and multilingual graph-building techniques to tackle the challenges of Wikipedia's scale, structure, and inherent noise. With a modular architecture and a simple graphical user interface (GUI), it is suitable for both technical and non-technical users. Built for scalability and reproducibility, WikiTextGraph supports interdisciplinary research in network science, computational linguistics, and digital humanities. Its flexible design enables easy adaptation for tasks involving low-resource or cross-lingual language studies.[1]

CORRESPONDING AUTHOR:
Gustavo Ariel Schwartz
Centro de Física de Materiales (CSIC – UPV/EHU), San Sebastián, Spain
gustavo.schwartz@csic.es

KEYWORDS:
Wikipedia; data mining; complex networks; computational language analysis; cultural analytics

TO CITE THIS ARTICLE:
Agapitos P, Suárez J-L, Schwartz GA 2025 WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction. *Journal of Open Research Software*, 13: 17. DOI: https://doi.org/10.5334/jors.572

(1) OVERVIEW

INTRODUCTION

Wikipedia[2] is one of the most frequently visited websites in the world and the most extensive online encyclopaedia. As of December 2024, it ranked among the top ten most popular websites globally, occupying 5th place according to Semrush[3] and 7th place according to Similarweb.[4] Internal statistics from the February 2025 report approximately 6,956,076 content pages and 48,729,378 registered users on the English version alone.[5] Due to its decentralized organization and independently managed communities, Wikipedia has attracted significant attention in academic research [1–3], serving as a valuable resource for studies on knowledge representation and information dissemination.

Extensive research has leveraged Wikipedia for diverse applications, including training neural networks [4–6], extracting structured data [7, 8], and conducting computational social science investigations. In this context, Wikipedia has been used for bias detection and large-scale analyses of biographies to explore cultural evolution [9–12]. Additionally, it plays a central role in Natural Language Processing (NLP) research, serving as a training corpus for models such as *MAVEN*, *BERT*, and *Wikiformer* [13–15].

Beyond textual analysis, Wikipedia's hyperlink structure has inspired numerous graph-based studies. Consonni *et al*. [16] released a multilingual link network dataset covering nine languages, with data snapshots spanning the years 2001 to 2018. Aspert *et al*. [17] developed a graph-structured dataset that integrates spatio-temporal data, although it is limited to the English edition. The YAGO project, including YAGO2 and YAGO3, introduced a knowledge base that enriches Wikipedia with disambiguation and temporal/spatial metadata, enhancing its semantic structure across multiple languages [18–20]. Similarly, Wu *et al*. [21] proposed a framework for constructing knowledge graphs from online encyclopaedias using heuristic matching and semi-supervised learning.

In addition to the previous, Wikipedia has proven to be a valuable resource for research on semantic relatedness and knowledge representation. For instance, Yeh *et al*. [22] measured semantic similarity between concepts using Wikipedia's content. More recently, Wang *et al*. [23] introduced WikiGraphs, a dataset that pairs graph structures with textual data to support tasks such as conditional text generation and representation learning. Similarly, Arroyo-Machado *et al*. [24] developed *Wikinformetrics*, a richly structured English Wikipedia dataset that links articles, categories, authors, external links, and references—providing a valuable foundation for informetric and bibliometric studies. Complementing this, Yang and Colavizza [25] investigated Wikipedia's growing use of academic citations, a trend Lewowieski [26] explored across 309 language editions, revealing significant differences in citation behaviour across cultures.

Moreover, several studies have explicitly focused on the link network structure of Wikipedia. Works such as Jatowt *et al*. [11], Consonni *et al*. [16], and Gabella [27] emphasise the need to resolve redirects to their final target pages to ensure accurate network representations—an essential preprocessing step before any formal analysis.

In the field of historical and cultural network analysis, high-quality Wikipedia graph data has been used to uncover latent patterns and relationships. Studies by Schwartz [28] and Miccio *et al*. [29, 30] demonstrate the value of clean and structured Wikipedia datasets for exploring cultural narratives and historiographical trends. This body of work highlights the utility of Wikipedia as an open data source and the importance of robust, reproducible methods for extracting and refining its complex graph structure.

Despite the substantial work in this area, effectively utilising Wikipedia's vast and dynamic content requires efficient extraction and processing tools. Several tools have been developed to address this challenge. "WikiExtractor" and "Cirrus Extractor" are Python-based tools designed to extract and clean plain text from Wikipedia XML dumps for applications in text analysis, graph construction, and machine learning [31].[6] The "wiki-dump-parser," implemented in Java, converts Wikipedia XML dumps into a structured JSON format, relying on "WikiExtractor" for text cleaning [32].[7] Another alternative is "Pywikibot" [33], a Python library that interacts directly with Wikipedia's MediaWiki API, enabling real-time data retrieval and redirect resolution but at the cost of significantly slower processing compared to raw dump extraction.

While existing tools and research have significantly advanced Wikipedia-based data extraction, they still exhibit critical gaps: some cannot parse data efficiently, others focus solely on a single language edition [17, 24] or rely heavily on external databases for multilingual integration [18–20], and many are not user-friendly enough to accommodate researchers with varied technical expertise [31–32]. WikiTextGraph addresses these shortcomings by providing a unified, scalable, and user-friendly solution that combines essential functionalities—parsing, text cleaning, graph extraction, redirect resolution, and broken link handling—into one coherent framework. This comprehensive approach enhances data quality for language modelling and refines graph-based analyses by eliminating distortions caused by unresolved redirects and non-content pages. Furthermore, WikiTextGraph is optimised for speed, memory efficiency, and multilingual processing, making it valuable for both high- and low-resource languages. For example, in about seven hours, it can complete extraction, cleaning, and graph construction for the English Wikipedia dump (January 2025). By consolidating these capabilities, WikiTextGraph streamlines workflows and offers a versatile, adaptable platform for large-scale Wikipedia research.

Following Wikipedia's open-edit philosophy, WikiTextGraph is openly available, allowing researchers

to contribute improvements or extend language support via pull requests. The tool is accessible on GitHub.[8]

IMPLEMENTATION AND ARCHITECTURE

WikiTextGraph is designed to efficiently process large-scale Wikipedia XML dumps, extract relevant textual content, and—if prompted—construct graph representations of Wikipedia's internal link network. The software follows a modular architecture that enables scalability, efficient memory management, multilingual support and extensibility. This section outlines the general structure of the software's workflow and its key components.

At a high level, the WikiTextGraph workflow consists of two main stages (see Figure 1):

1. **Parsing and Cleaning:** Wikipedia XML dumps are incrementally processed to extract relevant textual content, while filtering out non-content pages and removing non-content sections, infoboxes, and other irrelevant elements.
2. **Graph Construction (optional):** The extracted links between Wikipedia articles are structured into a directed graph, ensuring data integrity through deduplication, broken link removal, and redirect resolution.

COMMAND-LINE INTERFACE (CLI)

Advanced users can configure the software through various parameters such as the Wikipedia dump file path, language code, and output directory:

```
python wikitextgraph.py
        --dump_filepath /path/to/dump.bz2 \
        --language_code en \
        --base_dir /output/directory \
        --generate_graph
```

GRAPHICAL USER INTERFACE (GUI)

A GUI is also available to provide an interactive workflow for users unfamiliar with command-line operations, guiding them through selecting input files, language preferences, and whether to generate a graph (see Figure 2). After clicking the "Start Processing" button, the GUI closes to free up some memory. The GUI can be used again by running the program anew.

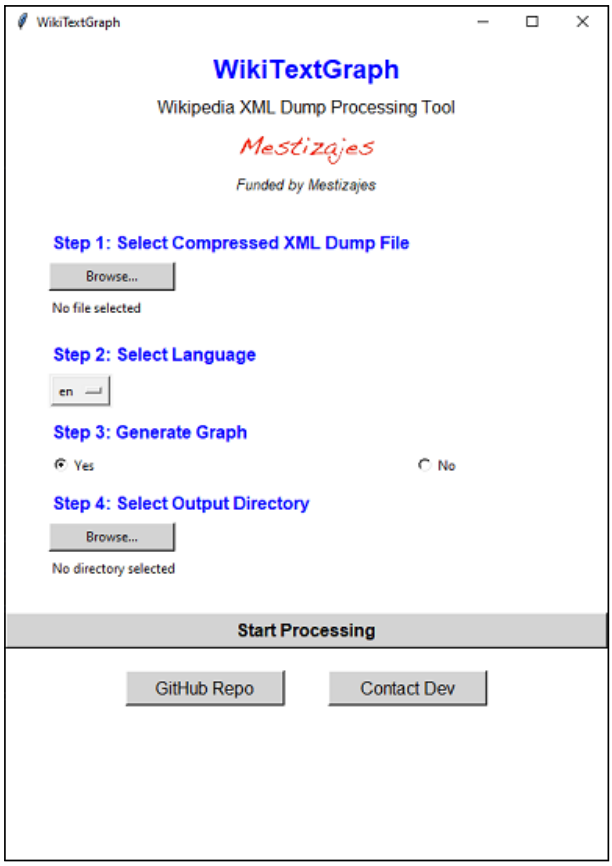

Figure 2 The Graphic User Interface (GUI) of the WikiTextGraph algorithm.

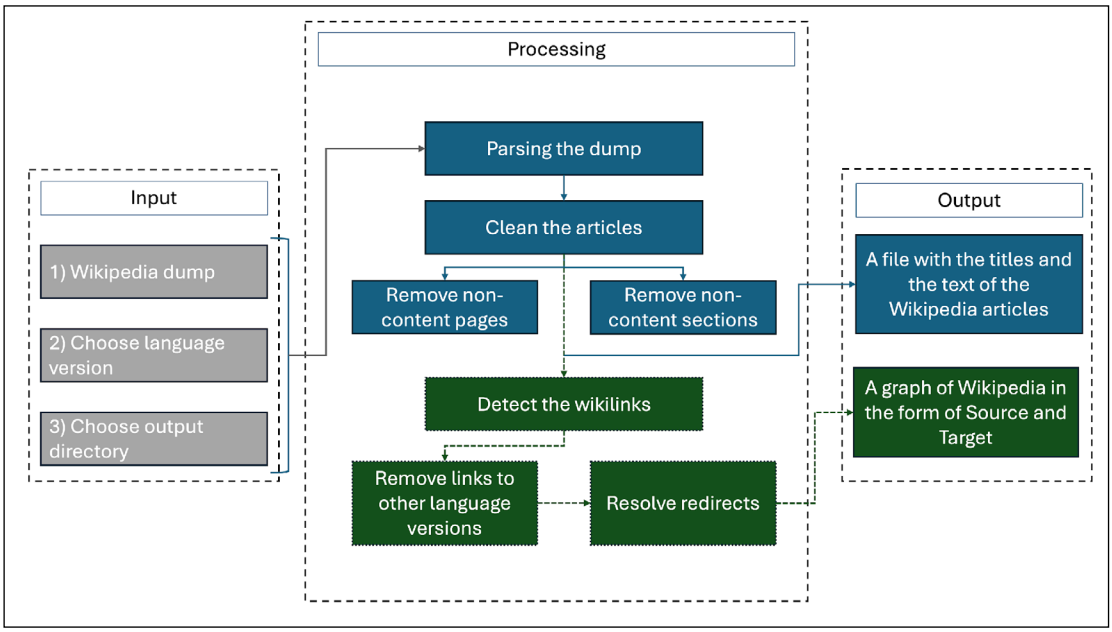

Figure 1 Workflow of how the software parses and, if prompted, generates the graph.

INSTALLATION AND EXECUTION

1. Clone the Repository
   `git clone https://github.com/PaschalisAg/WikiTextGraph`
2. Install Dependencies
   `pip install -r requirements.txt`
3. Run the Software
   `python wikitextgraph.py`

DATA MANAGEMENT AND PROCESSING

This section describes how each stage in WikiTextGraph's pipeline is carried out, explaining the motivations for each filtering strategy and data-handling step. Figure 1 provides a visual overview of the process.

INPUT

WikiTextGraph requires the following inputs to execute its pipeline:

1. A compressed Wikipedia XML dump file (e.g., `.bz2`) corresponding to one of the supported language editions.[9]
2. A language code indicating which Wikipedia language version to process (e.g., `en` for English, `es` for Spanish, etc.).[10]
3. A valid path to the output directory where the cleaned text files and, optionally, the graph files will be stored.

PARSING AND FILTERING WIKIPEDIA ARTICLES

The `parser_module` parses the compressed Wikipedia XML dump in batches (by default, 10,000 articles per batch) using a SAX-based approach[11] and systematically extracts the core content from each article while excluding non-informative pages (see Table 1) and sections (see Table 2). This step ensures that for each content page, only the textual content of each article is retained for subsequent analyses. After processing each batch, the system writes the cleaned content to disk immediately to reduce memory overhead.

REMOVING IRRELEVANT CONTENT

Pages for administrative or bibliographical purposes are identified through regular-expression checks (e.g., Draft:, Template:) and omitted from further processing (see Table 1). Configurable patterns in `LANG_SETTINGS` can be refined for additional languages or unique research needs.

PRODUCING CLEANED AND STRUCTURED TEXT

After filtering pages, WikiTextGraph applies cleaning routines to remove HTML comments, `<ref>` tags and extraneous wikitext templates that may exist inside an article. Sections such as References or External links are also discarded (see Table 2). The result is a structured, cleaned version of the main body for each article, ready for text-based analyses.

BUILDING THE DIRECTED GRAPH

If a user requests graph construction (through the CLI or GUI), the graph module creates a Source-Target dataset for each internal link, producing a directed network suitable for graph-based research. Node names are saved in a numerical format (i.e., each node is assigned a unique numerical ID), allowing faster loading and processing of large graphs later.

EXTRACTING INTERNAL LINKS

Internal links are identified using a regular expression that captures various wiki markup patterns (e.g., sections, alternative text) while preserving only the main article link. The regular expression used is shown below:

```
r'\[\['
r'([^#|]+)'
r'(?:#([^|]]))?'
r'(?:\|([^]]*))?'
r'\]\]'
```

The following example (dummy text generated by ChatGPT) demonstrates how various patterns of wikilinks can appear in an article and what the regular expression is going to capture (text in bold):

```
Quantum dots are [[semiconductor]]
nanocrystals that exhibit [[quantum
confinement|unique optical properties]].
These structures can be used in various
applications, including [[display
technology#Quantum dots|advanced display
technology]] and [[solar cells#Efficiency]]
research. Unlike traditional materials,
[[quantum dots#Excitonic behavior|their
excitonic behavior]] allows for highly
tunable optical properties. The development
of [[nanotechnology]] has accelerated
research in this field.
```

ENSURING GRAPH QUALITY

The algorithm begins by correcting malformed links through normalisation, for example, converting underscores to spaces (e.g., for instance, `Murray_Bookchin` will be converted to `Murray Bookchin` in order to match the actual Wikipedia page). Redirects are resolved by mapping them to their corresponding target, ensuring uniformity across page references in an article. To preserve the integrity of the dataset, links to other language editions are excluded, thereby limiting the network to a single language version (see Table 3).

ENGLISH (en)				
Wiktionary:	Category:	Draft:	File:	List of
MediaWiki:	Module:	Template:	Wikipedia:	Index of
Help:	Portal:	Image:	(disambiguation)	
SPANISH (es)				
Wiktionary:	Categoría:	File:	Archivo:	Image:
MediaWiki:	Plantilla:	Wikipedia:	Anexo:	Módulo:
Portal:	Help:	Ayuda:	Wikiproyecto:	Usuario:
User:		(desambiguación)		
GREEK (el)				
Wiktionary:	Κατηγορία:	Αρχείο:	File:	Image:
MediaWiki:	Module:	Πρότυπο:	Wikipedia:	Βικιπαίδεια:
(αποσαφήνιση)	Portal:	Πύλη:	Βοήθεια:	Topic:
Χρήστης:	User:			
POLISH (pl)				
Wikipedia:	Pomoc:	Szablon:	MediaWiki:	Kategoria:
Wikiprojekt:	Portal:	Plik:	Moduł:	User:
Wątek:	Topic:	(ujednoznacznienie)		
ITALIAN (it)				
Wikipedia:	Aiuto:	Template:	MediaWiki:	Categoria:
Progetto:	Portale:	File:	Modulo:	Topic:
(disambigua)	(disambiguazione)			
DUTCH (nl)				
Wikipedia:	Help:	Sjabloon:	MediaWiki:	Categorie:
Portaal:	Bestand:	Module:	(disambiguatie:)	(disambiguation)
User:				
BASQUE (eu)				
Wikipedia:	Laguntza:	Txantilloi:	MediaWiki:	Kategoria:
Maila:	Atari:	Ataria:	Usuario:	Modulu:
Fitxategi:	Wikiproiektua:	Eranskina:	Txikipedia:	Zerrenda:
(argipena)				
HINDI (hi)				
विकिपीडिया:	साँचा:	श्रेणी:	मीडियाविकि:	सहायता:
रवेशद्वार:	चित्र:	विकिपरियोजना:	मॉड्यूल:	(बहुविकल्पी)
GERMAN (de)				
Wikipedia:	Hilfe:	Vorlage:	MediaWiki:	Kategorie:
Portal:	Benutzer:	Modul:	Datei:	Liste der
Liste des	Liste von	(begriffsklärung)		
VIETNAMESE (vi)				
Wikipedia:	MediaWiki:	Trợ giúp:	Bản mẫu:	Tập tin:
Thể loại:	Sách:	Danh sách:	Cổng thông tin:	Mô đun:
(định hướng)				

Table 1 Pages the algorithm removes during the text cleaning phase for each language version supported by the algorithm.

ENGLISH (en)			
See also	Publications	References	Notes
Footnotes	External links	Further Reading	Draft:
SPANISH (es)			
Véase también	Referencias	Notas	Enlaces externos
Bibliografía	Otra lectura		
GREEK (el)			
Δείτε επίσης	Παραπομπές	Σημειώσεις	Εξωτερικοί σύνδεσμοι
Προτεινόμενη βιβλιογραφία	Βιβλιογραφία	Περαιτέρω ανάγνωση	
POLISH (pl)			
Zobacz też	Uwagi	Bibliografia	Przypisy
Linki zewnętrzne	Literatura w języku polskim		
ITALIAN (it)			
Note	Bibliografia	Altri progetti	Collegamenti esterni
Riferimenti	Voci correlate		
DUTCH (nl)			
Noten	Bibliografie	Appendix	Externe links
Referenties	Bronnen	Zie ook	Overig
Ander	Literatuur	Voetnoten	
BASQUE (eu)			
Oharrak	Bibliografia	Ahultasunak	Kanpo estekak
Erreferetziak	Ikus, gainera		
HINDI (hi)			
ण्णिियांट	यह भी देखिए	संदर्भ सूची	बाहरी कड़ियाँ
सन्दर्भ	इन्हें भी देखें	इसके अतिरिक्त पठन	
GERMAN (de)			
Weblinks	Einzelnachweise	Einzelnachweise und Anmerkungen	Anmerkungen
Literatur	Siehe auch	Fußnoten	Veröffentlichungen
VIETNAMESE (vi)			
Xem thêm	Tham khảo	Liên kết ngoài	Tài liệu tham khảo
Tài liệu	Ghi chú	Chú thích	Tài liệu khác
Hình ảnh	Đọc thêm		

Table 2 Sections where extracting text stops. If one of these sections appears in the article, it indicates the stopping point for extracting content.

Additionally, self-loops and duplicate edges are removed to eliminate redundancy. Together, these validation steps produce a cleaner and more coherent graph that more accurately reflects meaningful inter-article connections. By filtering out structural noise introduced by non-existent connections and navigational features, such as redirects and interlanguage links, the resulting topology is better suited for downstream tasks, including community detection.

PARQUET FORMAT FOR GRAPH DATA

The resulting source-target pairs are stored in parquet files in a numerical format, offering advantages in compression and read/write and space-storing efficiency. This columnar format permits partial column reading during queries or large-scale analyses, making it well-suited to network exploration tasks.

COMPRESSED REDIRECT MAPPINGS

WikiTextGraph stores redirect information in a compressed pickle file, mapping each redirect page to its canonical target in an inverse-dictionary format. This design minimises disk usage and lookup times, allowing researchers to analyse redirect patterns independently and enabling near-instant lookups during graph construction.

ENGLISH (en)			
#redirect			
SPANISH (es)			
#redirección		#redirect	
GREEK (el)			
#ανακατεύθυνση		#redirect	
POLISH (pl)			
#patrz	#redirect	#przekieruj	#tam
ITALIAN (it)			
#rinvia		#redirect	
DUTCH (nl)			
#doorverwijzing		#redirect	
BASQUE (eu)			
#birzuzendu		#redirect	
HINDI (hi)			
#अनुप्रेषित	#पुनर्प्रेषित	#redirect	
GERMAN (de)			
#weiterleitung		#redirect	
VIETNAMESE (vi)			
#đổi		#redirect	

Table 3 Table with the keywords to detect redirects for each language version.

FINAL OUTPUT

The software produces two main directories:

1. output: Contains two columns, one for the article titles and the other for the cleaned text.
2. graph: Stores the constructed graph with nodes represented by numerical IDs for fast loading and analysis (two columns in the format of Source and Target). This directory includes a lookup table for redirects-to-target and an ID-to-article's title mapping.

QUALITY CONTROL

Ensuring the reliability and accuracy of WikiTextGraph is very important, particularly in the absence of an official validation pipeline. We design and implement a comprehensive quality control framework consisting of multiple validation steps to achieve this. These checks assess the consistency of the extracted data at each stage of the extraction process across all supported language versions, ensuring alignment with the intended final output. The validation steps are as follows:

1. **Language Evaluation:** To assess the effectiveness of the algorithm in different language versions, we conduct an evaluation with native (L1) speakers. This step verifies the algorithm's ability to accurately capture language-specific patterns and nuances.
2. **Text Extraction and Filtering:** We verify that predefined sections, such as references and external links, are correctly removed during the extraction process. Additionally, we ensure that non-content pages are effectively excluded from the dataset as defined by our filtering criteria. Our evaluation confirms that the algorithm performs as expected in both cases.
3. **Edge Validity Check:** We validate that all extracted edges correspond strictly to valid Wikipedia pages. This step eliminates invalid or extraneous links, including self-loops, broken links, and links pointing to specific sections within pages, thereby maintaining the structural integrity of the constructed graph.
4. **Node Count Verification:** To confirm completeness, we compare the number of nodes in the generated graphs with the official count of content pages reported by Wikipedia for each language version. While the observed node count is slightly lower than the reported numbers, this discrepancy is expected due to the deliberate removal of certain pages (see Table 4).
5. **Graph Structural Analysis:** To validate the network structure, we analyse the degree distributions of the constructed graphs. The results confirm that these distributions exhibit heavy-tailed, power-law-like behaviour, a characteristic feature of Wikipedia's hyperlink network (see Figures 3, 4, and Table 4) [27, 34].
6. **Redirects and Interlanguage Link Validation:** We manually inspect randomly sampled entries during the graph construction phase to ensure correct link resolution. This evaluation confirms that all interlanguage links have been successfully removed and that all redirects are correctly replaced with their corresponding target pages.

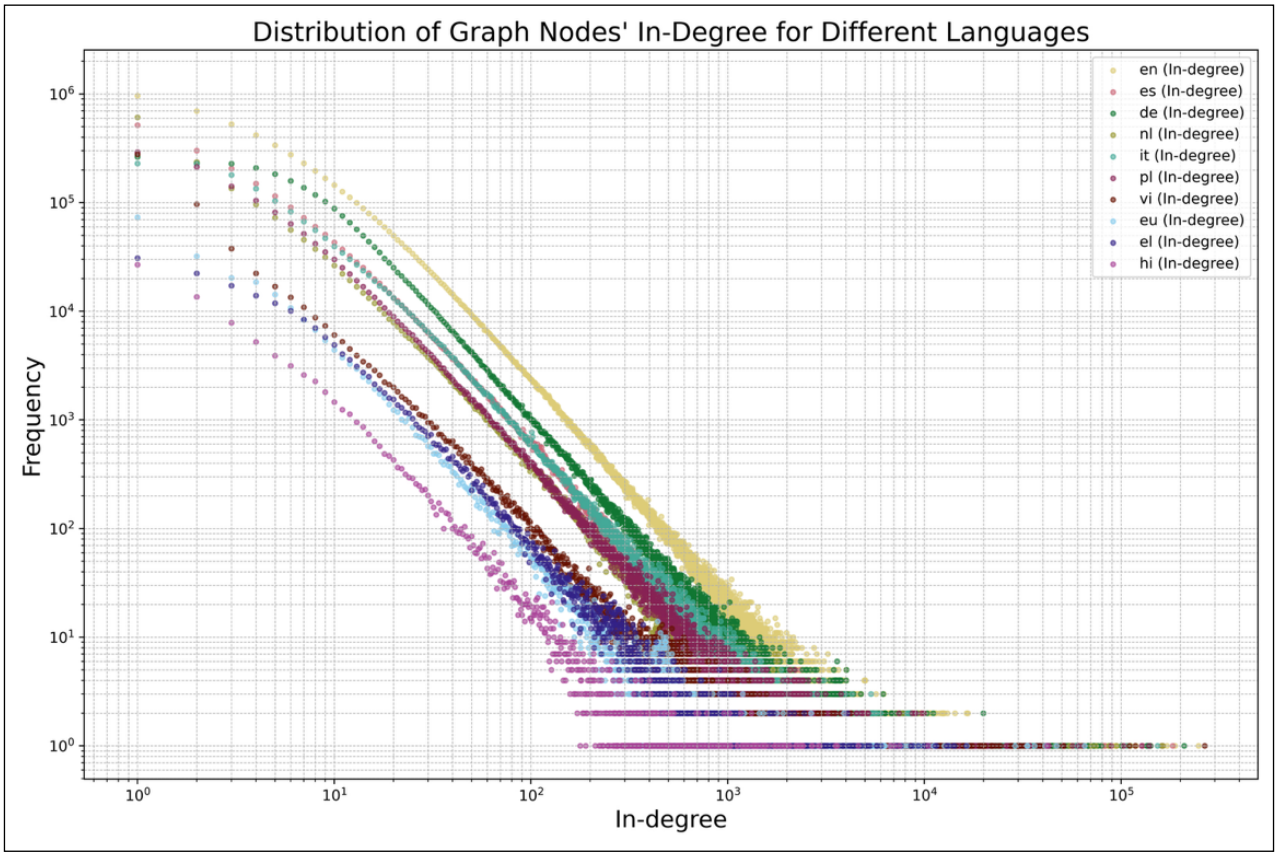

Figure 3 Log-log plot of the in-degree[12] distribution for each language version. The x-axis represents the in-degree (i.e., the number of incoming links to a node), whereas the y-axis represents the frequency of nodes for each in-degree value. Each curve corresponds to a different language version, as indicated in the legend.

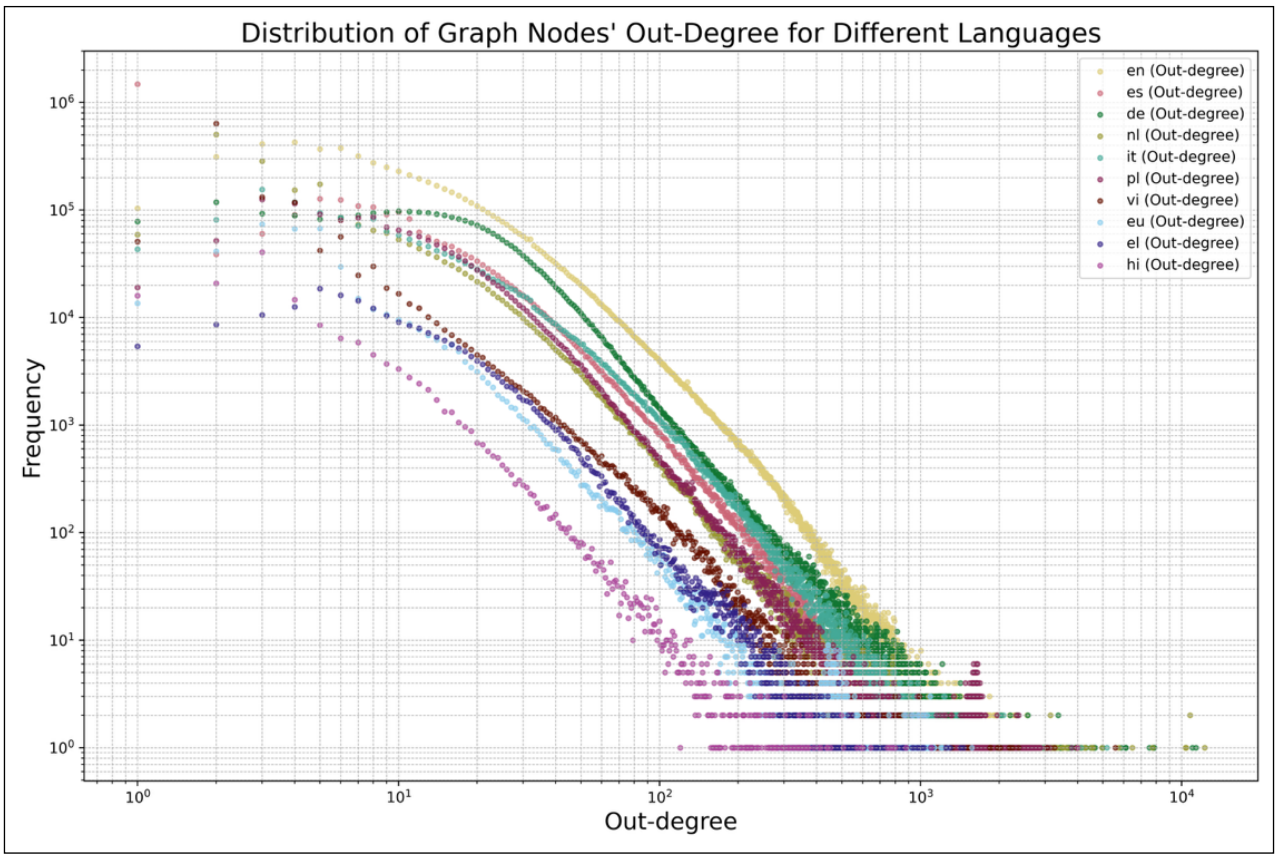

Figure 4 Log-log plot of the out-degree[13] distribution for each language version. The x-axis represents the out-degree (i.e., the number of outgoing links from a node), whereas the y-axis represents the frequency of nodes for each out-degree. Each curve corresponds to a different language version, as indicated in the legend.

LANGUAGE	ORDER	SIZE	AVERAGE IN/OUT DEGREE	MAX IN-DEGREE	MAX OUT-DEGREE	DENSITY ($\times 10^{-6}$)	DATE
English (en)	6,736,622	159,598,688	23.6	247,992	4,615	4	Jan. 2025
German (de)	3,001,074	77,213,282	25.7	209,493	11,359	9	Nov. 2024
Dutch (nl)	2,169,658	27,487,945	12.6	163,525	12,263	6	Nov. 2024
Spanish (es)	1,905,582	39,473,628	20.7	205,662	4,083	11	Nov. 2024
Italian (it)	1,820,884	42,649,948	23.4	153,337	5,179	13	Nov. 2024
Polish (pl)	1,620,570	30,140,616	18.5	138,619	3,800	11	Nov. 2024
Vietnamese (vi)	1,284,928	8,977,353	6.9	266,137	5,576	5	Jan. 2025
Basque (eu)	435, 429	4,436,022	10.1	65,339	1,573	23	Dec. 2024
Greek (el)	237,016	4,319,401	18.2	16,314	1,464	77	Nov. 2024
Hindi (hi)	152,990	1,097,849	7.1	42,123	3,993	47	Nov. 2024

Table 4 Summary statistics of the Wikipedia language versions supported (at the time of writing) and processed by WikiTextGraph. Each row corresponds to a different language version, identified by its name and ISO 639 language code in parentheses. The columns report the following metrics calculated using the Python library NetworkX [35]. *Order* is the number of nodes (Wikipedia articles) in the directed network; *Size* is the number of edges (links between articles); *Average In/Out degree* is the average number of incoming and outgoing links per node; *Max In/Out-degree* is the highest number of incoming and outgoing links respectively; *Density* is the ratio of actual links to all possible links[14]; *Date* is the month and the year that corresponds to the version of the Wikipedia dump collected and processed.

(2) AVAILABILITY

OPERATING SYSTEM

Linux (tested: Mint 21.3), macOS (tested: macOS 12 and newer), Windows (tested: Windows 10 and newer).

PROGRAMMING LANGUAGE

Python versions 3.9 and newer.

ADDITIONAL SYSTEM REQUIREMENTS

The software was developed and primarily tested in the following environment:

- **Operating System:** Linux Mint 21.3
- **Kernel:** Version 5.15
- **Processor:** Intel Core i7 (4 cores, 8 threads, 3.6 GHz)
- **Memory:** Minimum 8 GB RAM
- **Disk:** At least 15 GB of free storage
- **Swap**[15]: Minimum 2 GB
- **Graphics:** Integrated GPU (e.g., Intel HD Graphics 630)

The software is cross-platform and has been tested on macOS (Monterey and newer) and Windows 10. Disk usage depends on the specific Wikipedia edition processed. For reference, processing the English Wikipedia snapshot dated 2025-01-23 requires a minimum of 36 GB of free disk space.

DEPENDENCIES

- `click` == 8.1.8
- `cloudpickle` == 3.1.1
- `colorama` == 0.4.6
- `contourpy` == 1.0.1
- `cramjam` == 2.9.1
- `dask` == 2023.5.0
- `fastparquet` == 2024.2.0
- `fonttools` == 4.22.0
- `fsspec` == 2025.2.0
- `importlib_metadata` == 8.5.0
- `kiwisolver` == 1.3.1
- `locket` == 1.0.0
- `numpy` == 1.24.4
- `packaging` == 24.2
- `pandas` == 2.0.3
- `partd` == 1.4.1
- `Pillow` == 10.3.0
- `pyarrow` == 17.0.0
- `pyparsing` == 3.1.2
- `PySide6` == 6.6.3.1
- `PySide6_Addons` == 6.6.3.1
- `PySide6_Essentials` == 6.6.3.1
- `python-dateutil` == 2.9.0.post0
- `pytz` == 2025.1
- `PyYAML` == 6.0.2
- `regex` == 2024.11.6
- `shiboken6` == 6.6.3.1
- `six` == 1.16.0
- `toolz` == 1.0.0
- `tqdm` == 4.67.1
- `tzdata` == 2025.1
- `wcwidth` == 0.2.13
- `wikitextparser` == 0.56.3
- `zipp` == 3.20.2
- `tkinter` == 8.5

SOFTWARE LOCATION

Code repository

Name: *GitHub*

Identifier: https://github.com/PaschalisAg/WikiTextGraph

Licence: Apache License Version 2.0, January 2004

Date published: 02/03/2025

Archive

Name: *Zenodo*
Identifier: https://zenodo.org/records/16260544
Publisher: Paschalis Agapitos
Version: v1.0.0
Licence: Apache License Version 2.0, January 2004
Date published: 21/07/2025

LANGUAGE
English

(3) REUSE POTENTIAL

WikiTextGraph's robust design and flexible functionality make it inherently reusable across a broad spectrum of research and application domains, including Natural Language Processing (NLP), Data Mining, Graph-based studies, and Computational or Digital Humanities. By parsing, filtering, and analysing structured Wikipedia text—and optionally generating a graph representation—it supports an array of data-centric tasks such as automated content extraction, entity recognition, and information retrieval.

In network science, WikiTextGraph helps construct and examine knowledge graphs and graph-based databases, revealing hidden relationships among entities and enabling advanced semantic search. Incorporating pandas, pyarrow, and the parquet format readily handles large-scale datasets, ensuring efficient big data analytics and providing a foundation for machine learning workflows. The graph-first paradigm further facilitates social network analysis, clarifying complex interconnections among concepts. Researchers in computational linguistics and multilingual text analysis also benefit from its modular design, which enables straightforward cross-lingual content processing and contributes to projects focused on low-resource languages.

Moreover, WikiTextGraph's modular architecture underscores its adaptability. Language-specific filtering and parsing rules within the `LANG_SETTINGS.yml` file can be tailored for individual research agendas, allowing domain- or language-specific customization. Its code is cleanly split into distinct components—`graph.py` for graph structures, `gui.py` for the user interface, `parser_module.py` for parsing logic, and `utils.py` for support functions—making it more straightforward to either extend existing features or embed the software within other analytic pipelines. This clear structure promotes code reuse and effortless scalability, effectively lowering the barrier for researchers to integrate WikiTextGraph into their projects.

Finally, the software's open availability and support ecosystem further amplify its reusability. Users can contribute bug reports or improvements through GitHub pull requests or issues, while direct assistance is accessible via the "Contact us" button in the GUI (refer to Figure 2). Through this collaborative environment, we want to ensure that WikiTextGraph remains a versatile and evolving tool for the research community.

DATA ACCESSIBILITY STATEMENT

All code created and used in this research is available at: https://github.com/PaschalisAg/WikiTextGraph. It has been archived and is persistently available at: https://zenodo.org/records/16260544.

NOTES

1 https://github.com/PaschalisAg/WikiTextGraph.
2 https://www.wikipedia.org/.
3 https://www.semrush.com/website/top/.
4 https://www.similarweb.com/top-websites/.
5 https://en.wikipedia.org/wiki/Special:Statistics.
6 https://github.com/attardi/wikiextractor.
7 https://github.com/studerw/wiki-dump-parser.
8 https://github.com/PaschalisAg/WikiTextGraph.
9 A Wikipedia dump that contains all the articles for a given language version can be found at https://dumps.wikimedia.org/backup-index.html. Users should navigate to the section "Sql/XML dumps issues" to find entries corresponding to different language versions listed in the format `{language_code}wiki: Dump complete`. Selecting one of these entries leads to a list of downloadable files. The relevant file follows the naming convention `{language_code}wiki-{yyyymmdd}-pages-articles-multistream.xml.bz2`. This file serves as the input to the software and should be downloaded and saved locally on the user's computer.
10 All language codes follow the ISO 639 standardized nomenclature.
11 SAX is an event-driven XML parsing method that allows efficient, memory-light processing of large files by reading them sequentially without loading the entire document into memory.
12 The in-degree of a node refers to the number of incoming links directed towards that node.
13 The out-degree of a node refers to the number of outgoing links leaving from that node.
14 For directed graphs the density is given by $m/[n(n-1)]$, where n is the number of nodes and m is the number of edges.
15 Reserved portion of disk space used as virtual memory by the operating system when then physical RAM is fully utilized.

ETHICS AND CONSENT

The article does not contain any studies with human participants performed by any of the authors.

ACKNOWLEDGEMENTS

We extend our sincere gratitude to the native (L1) speakers who contributed their time and expertise in evaluating the results for each language version during the development of WikiTextGraph. Following the order in which the languages appear in the GUI, we would like to acknowledge the following individuals for their support formally:

- For Spanish (es) and Basque (eu): Amaia Elizaran Mendarte and Ane Escobar Fernández
- For Polish (pl): Adam Olejniczak and Zuzanna Lawera
- For Italian (it): Valerio Di Lisio
- For Hindi (hi): Anish Rao
- For German (de): Balthasar Braunewell
- For Vietnamese (vi): Phuong Thu Le

Their contributions were significant in ensuring the model's quality and linguistic accuracy across multiple language settings.

FUNDING INFORMATION

GAS acknowledges the financial support from the Spanish Government, project PID2023-146348NB-I00 financed by MICIU/AEI/10.13039/501100011033 and by FEDER (UE) and the Basque Government (IT-1566-22). PA and GAS acknowledge the financial support from the Donostia International Physics Centre (Programa Mestizajes) and the support of NVIDIA Corporation, which provided a donation of a Quadro RTX 6000 GPU.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Paschalis Agapitos software development, data analysis, conceptualisation, and writing, including original draft preparation, review, and editing.
Juan Luis Suárez contributed to writing through review and editing.
Gustavo Ariel Schwartz conceptualisation, data analysis, software review, writing (original draft, review, and editing), and secured funding for the project.

AUTHOR AFFILIATIONS

Paschalis Agapitos orcid.org/0000-0001-6809-3635
Donostia International Physics Center (DIPC), San Sebastián, Spain; Centro de Física de Materiales (CSIC – UPV/EHU), San Sebastián, Spain; Department of Philosophy, University of the Basque Country (UPV/EHU), Spain
Juan-Luis Suárez orcid.org/0000-0001-9302-2986
CulturePlex Lab, Western University, London, Ontario, Canada
Gustavo Ariel Schwartz orcid.org/0000-0003-3044-2435
Centro de Física de Materiales (CSIC – UPV/EHU), San Sebastián, Spain

REFERENCES

1. **Callahan ES, Herring SC.** Cultural bias in Wikipedia content on famous persons. *J Am Soc Inf Sci Technol*. 2011;62(10):1899–1915. DOI: https://doi.org/10.1002/asi.21577
2. **Park TK.** The visibility of Wikipedia in scholarly publications. *First Monday*. Published online 2011. DOI: https://doi.org/10.5210/fm.v16i8.3492
3. **Areia C, Burton K, Taylor M, Watkinson C.** Research citations building trust in Wikipedia: Results from a survey of published authors. *PLOS One*. 2025;20(4):e0320334. DOI: https://doi.org/10.1371/journal.pone.0320334
4. **Fan A, Gardent C.** Generating Biographies on Wikipedia: The Impact of Gender Bias on the Retrieval-Based Generation of Women Biographies. *Proc 60th Annu Meet Assoc Comput Linguistics (Vol 1: Long Pap)*. Published online 2022;8561–8576. DOI: https://doi.org/10.18653/v1/2022.acl-long.586
5. **Xu S, Liu S, Culhane T, Pertseva E, Wu MH, Semnani S, Lam M.** Fine-tuned LLMs Know More, Hallucinate Less with Few-Shot Sequence-to-Sequence Semantic Parsing over Wikidata. *Proc 2023 Conf Empir Methods Nat Lang Process*. Published online 2023;5778–5791. DOI: https://doi.org/10.18653/v1/2023.emnlp-main.353
6. **Peng Y, Bonald T, Alam M.** Refining Wikidata Taxonomy using Large Language Models. *arXiv*. Published online 2024. DOI: https://doi.org/10.1145/3627673.3679156
7. **Bhole A, Fortuna B, Grobelnik M, Mladenic D.** Extracting named entities and relating them over time based on Wikipedia. *Informatica*. 2007;31(4). https://www.informatica.si/index.php/informatica/article/view/169
8. **BØhn C, Nørvåg K.** Extracting Named Entities and Synonyms from Wikipedia. *2010 24th IEEE Int Conf Adv Inf Netw Appl*. Published online 2010;1300–1307. DOI: https://doi.org/10.1109/AINA.2010.50
9. **Riehle D, Gonzalez-Barahona JM, Robles G, Möslein KM, Schieferdecker I, Cress U, Wichmann A, Hecht B, Jullien N, Viseur R.** Reliability of User-Generated Data. *Proc Int Symp Open Collab*. Published online 2014;1–3. DOI: https://doi.org/10.1145/2641580.2641618
10. **Yu AZ, Ronen S, Hu K, Lu T, Hidalgo CA.** Pantheon 1.0, a manually verified dataset of globally famous biographies. *Sci Data*. 2016;3(1):150075. DOI: https://doi.org/10.1038/sdata.2015.75
11. **Jatowt A, Kawai D, Tanaka K.** Digital History Meets Wikipedia. *Proc 16th ACMIEEE-CS Jt Conf Digit Libr*. Published online 2016;17–26. DOI: https://doi.org/10.1145/2910896.2910911
12. **Beytía P, Schobin J.** Networked Pantheon: a Relational Database of Globally Famous People: Social and Behavioural Sciences. *Res Data J Humanit Soc Sci*. 2020;5(1):50–65. DOI: https://doi.org/10.1163/24523666-00501002
13. **Wang X, Wang Z, Han X, Jiang W, Han R, Liu Z, Li J, Li P, Lin Y, Zhou J.** MAVEN: A Massive General Domain Event

Detection Dataset. *Proc 2020 Conf Empir Methods Nat Lang Process (EMNLP)*. Published online 2020;1652–1671. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.129

14. **Liu Z, He X, Liu L, Liu T, Zhai X.** Context Matters: A Strategy to Pre-train Language Model for Science Education. *Commun Comput Inf Sci*. Published online 2023;666–674. DOI: https://doi.org/10.1007/978-3-031-36336-8_103
15. **Su W, Ai Q, Li X, Chen J, Liu Y, Wu X, Hou S.** Wikiformer: Pre-training with Structured Information of Wikipedia for Ad-hoc Retrieval. *arXiv*. Published online 2023. DOI: https://doi.org/10.1609/aaai.v38i17.29869
16. **Consonni C, Laniado D, Montresor A.** WikiLinkGraphs: A Complete, Longitudinal and Multi-Language Dataset of the Wikipedia Link Networks. *arXiv*. Published online 2019. DOI: https://doi.org/10.1609/icwsm.v13i01.3257
17. **Aspert N, Miz V, Ricaud B, Vandergheynst P.** A Graph-Structured Dataset for Wikipedia Research. *Companion Proc 2019 World Wide Web Conf*. Published online 2019;1188–1193. DOI: https://doi.org/10.1145/3308560.3316757
18. **Rebele T, Suchanek F, Hoffart J, Biega J, Kuzey E, Weikum G.** YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. *Lect Notes Comput Sci*. Published online 2016;177–185. DOI: https://doi.org/10.1007/978-3-319-46547-0_19
19. **Biega J, Kuzey E, Suchanek FM.** Inside YAGO2s. *Proc 22nd Int Conf World Wide Web*. Published online 2013;325–328. DOI: https://doi.org/10.1145/2487788.2487935
20. **Mahdisoltani F, Biega J, Suchanek FM.** YAGO3: A knowledge base from multilingual Wikipedias. Published 2013. https://imt.hal.science/hal-01699874
21. **Wu T, Wang H, Li C, Qi G, Niu X, Wang M, Li L, Shi C.** Knowledge graph construction from multiple online encyclopedias. *World Wide Web*. 2020;23(5):2671–2698. DOI: https://doi.org/10.1007/s11280-019-00719-4
22. **Yeh E, Ramage D, Manning CD, Agirre E, Soroa A.** WikiWalk: random walks on Wikipedia for semantic relatedness. In: *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics; 2009. DOI: https://doi.org/10.3115/1708124.1708133
23. **1. Wang L, Li Y, Aslan O, Vinyals O.** WikiGraphs: A Wikipedia Text – Knowledge Graph Paired Dataset. *arXiv*. Published online 2021. DOI: https://doi.org/10.18653/v1/2021.textgraphs-1.7
24. **Arroyo-Machado W, Torres-Salinas D, Costas R.** Wikinformetrics: Construction and description of an open Wikipedia knowledge graph data set for informetric purposes. *Quant Sci Stud*. 2022;3(4):931–952. DOI: https://doi.org/10.1162/qss_a_00226
25. **Yang P, Colavizza G.** A Map of Science in Wikipedia. *Companion Proc Web Conf 2022*. Published online 2022;1289–1300. DOI: https://doi.org/10.1145/3487553.3524925
26. **Lewoniewski W.** The Most Cited Scientific Information Sources in Wikipedia Articles Across Various Languages. *Biblioteka*. 2024;27(36):269–294. DOI: https://doi.org/10.14746/b.2023.27.12
27. **Gabella M.** Cultural Structures of Knowledge from Wikipedia Networks of First Links. *Ieee Transactions Netw Sci Eng*. 2017;6(3):249–252. DOI: https://doi.org/10.1109/TNSE.2018.2812788
28. **Schwartz GA.** Complex networks reveal emergent interdisciplinary knowledge in Wikipedia. *Humanit Soc Sci Commun*. 2021;8(1):127. DOI: https://doi.org/10.1057/s41599-021-00801-1
29. **Miccio LA, Gámez-Pérez C, Suárez JL, Schwartz GA.** Mapping the Networked Context of Copernicus, Michelangelo, and Della Mirandola in Wikipedia. Ramona R, Maximillian S, Hyejin Y, Mikhail T, editors *ACS*. 2022;25(05n06):2240010-1-2240010-2240012. DOI: https://doi.org/10.1142/S0219525922400100
30. **Miccio LA, Agapitos P, Gamez-Perez C, González F, Suarez JL, Schwartz GA.** Wikipedia as a cultural lens: a quantitative approach for exploring cultural networks. *Humanit Soc Sci Commun*. 2025;12(1):462. DOI: https://doi.org/10.1057/s41599-025-04772-5
31. **Attardi G.** attardi/wikiextractor; 2025. https://github.com/attardi/wikiextractor
32. wiki-dump-parser: A simple but fast Python script that reads the XML dump of a wiki and outputs the processed data in a CSV file. https://github.com/Grasia/wiki-scripts/tree/master/wiki_dump_parser
33. wikimedia/pywikibot 2025. https://github.com/wikimedia/pywikibot
34. **Voss J.** Measuring Wikipedia. In: *Proceedings of ISSI*. Vol 1. 2005 http://eprints.rclis.org/6207/
35. **Hagberg AA, Schult DA, Swart PJ.** Exploring Network Structure, Dynamics, and Function using NetworkX. *Proc 7th Python Sci Conf*. Published online 2008;11–15. DOI: https://doi.org/10.25080/TCWV9851

TO CITE THIS ARTICLE:
Agapitos P, Suárez J-L, Schwartz GA 2025 WikiTextGraph: A Python Tool for Parsing Multilingual Wikipedia Text and Graph Extraction. *Journal of Open Research Software*, 13: 17. DOI: https://doi.org/10.5334/jors.572

Submitted: 15 April 2025 **Accepted:** 28 July 2025 **Published:** 12 September 2025

COPYRIGHT:
© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Research Software is a peer-reviewed open access journal published by Ubiquity Press.