The Journal of Open Source Software

The Irchel Geoparser: A Modular Python Library for Toponym Recognition and Resolution

Diego Gomes [1¶] **and Ross S. Purves** [1]

1 Department of Geography, University of Zurich, Switzerland ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 28 November 2025
Published: unpublished

Summary

The Irchel Geoparser is an open-source Python library for recognising and resolving place names (toponyms) in unstructured text. Based on a modular architecture, it provides an end-to-end geoparsing pipeline in which the components responsible for toponym recognition, toponym resolution, and the underlying gazetteer data can be freely adapted or exchanged. The library includes robust, configurable built-in modules and a GeoNames-based gazetteer configuration, offering strong performance for English text out of the box. Users can also choose to fine-tune the built-in modules or integrate their own custom recognisers and resolvers. Similarly, it is possible to include domain-specific gazetteers, allowing geoparsing workflows to be tailored to diverse research and practical contexts.

Statement of Need

Geoparsing is an essential step in linking text to geographical locations (Purves et al., 2018), enabling diverse applications including mapping news coverage, exploring historical documents, or analysing discourse across regions (Chesnokova et al., 2019; Grover et al., 2010; Teitler et al., 2008). Although several geoparsers already exist, they have limitations that restrict their practical applicability. Systems such as the Edinburgh Geoparser (Grover et al., 2010), CLAVIN (Berico Technologies, 2012), GeoTxt (Karimzadeh et al., 2019), and the Mordecai family (Halterman, 2017, 2023) have contributed significantly to the field, but they generally require substantial effort for deployment and offer only limited flexibility in the customisation of the geoparsing process. Adapting these tools to a specific task, whether by incorporating new models for languages other than English, using specialised gazetteers or aligning the pipeline to a domain typically entails maintaining a separate fork of the software.

The Irchel Geoparser addresses these limitations by providing a clearly defined, interchangeable geoparsing pipeline packaged as an installable Python library. Its built-in modules offer strong default performance for English text and all stages of the pipeline can be adapted or exchanged entirely. This makes the Irchel Geoparser suitable both as a ready-to-use tool and as an experimental platform for researchers developing new geoparsing methods. By developing a modular architecture, we make it possible for other researchers to integrate customised modules addressing individual elements of the geoparsing workflow - for example recent work using LLMs to enhance toponym recognition (Yan & Lee, 2024) or resolution (Anuradha et al., 2025; Hu et al., 2024; Yan & Lee, 2024).

Architecture

The Irchel Geoparser treats geoparsing as a pipeline composed of recogniser and resolver modules that adhere to simple, well-defined interfaces. Any Python class implementing these interfaces can be used within the pipeline, allowing users to integrate custom methods while

relying on the system for data handling, caching, and tracking. In this structure, documents and geoparsing results are managed within a local SQLite database, enabling persistent storage for long-term workflows, by maintaining a project structure in which raw text, recognised toponyms, resolver outputs and auxiliary metadata are stored together.

The built-in recogniser uses spaCy's named-entity recognition functionality to identify toponyms based on entity labels such as Geopolitical Entity (GPE), Location (LOC), or Facility (FAC). Because it builds directly on spaCy models, users can easily switch languages, use fine-tuned models or incorporate custom pipelines by simply specifying a different model. The built-in resolver incorporates an embedding-based disambiguation method presented in prior work (Gomes et al., 2024). It retrieves candidate locations from a gazetteer, embeds both the mention context and candidate descriptions using a transformer model, and ranks candidates based on embedding similarity. Users can choose to use the default model that has been fine-tuned using English news articles linked to GeoNames, or they can supply their own transformer checkpoints, which in turn can be trained for use with any configured gazetteer.

The gazetteer layer itself is also customisable. While a GeoNames-based configuration is provided by default, users can create gazetteer configuration files to build custom gazetteers, combining different data sources as required. This makes integrating gazetteer data created for, for example, historical georeferencing (Grossner et al., 2023) or by national mapping agencies, straightforward. Since modules only depend on the gazetteer interface, switching between different gazetteers requires no additional code changes.

Applications

The Irchel Geoparser has been used in both research and teaching contexts. For example, it supported a study on geographic bias in global climate disaster reporting, in which thousands of news articles were analysed and mapped to reveal patterns in the geographic focus of media attention (Kong & Purves, 2025). It was also used in a Master's thesis that investigated the wolf debate in Switzerland, enabling shifts in the geographic focus of the debate to be mapped over time (Besse, 2025). Straightforward installation, comprehensive documentation, and a complete demo available as a Docker image also make the Irchel Geoparser an accessible tool for teaching and for illustrating key concepts in geographic information retrieval.

Acknowledgements

The development of the Irchel Geoparser was supported by the GIS Hub, a project funded by the Digital Society Initiative (DSI) of the University of Zurich. Additional funding was provided by the Public Data Lab of the Digitalization Initiative of the Zurich Higher Education Institutions (DIZH). We thank Michele Volpi for his involvement in developing the first version of the Irchel Geoparser, and Nicolas Spring for his contributions as an open-source collaborator.

References

Anuradha, I., Sumanathilaka, D. K., Mitkov, R., & Rayson, P. (2025). Toponym resolution: Will prompt engineering change expectations? *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing*, 95–104. https://acl-bg.org/proceedings/2025/RANLP%202025/pdf/2025.ranlp-1.11.pdf

Berico Technologies. (2012). *CLAVIN (Cartographic Location And Vicinity INdexer)*.

Besse, M. (2025). *The wolf debate in switzerland: A spatial and temporal analysis performing geographic information retrieval on newspaper articles* [Master's thesis]. University of Zurich.

Chesnokova, O., Taylor, J. E., Gregory, I. N., & Purves, R. S. (2019). Hearing the silence: Finding the middle ground in the spatial humanities? Extracting and comparing perceived silence and tranquillity in the english lake district. *International Journal of Geographical Information Science*, *33*(12), 2430–2454. https://doi.org/10.1080/13658816.2018.1552789

Gomes, D., Purves, R. S., & Volpi, M. (2024). Fine-tuning transformers for toponym resolution: A contextual embedding approach to candidate ranking. In X. Hu, R. S. Purves, L. Moncla, J. Kersten, & K. Stock (Eds.), *Proceedings of the GeoExT 2024: Geographic information extraction from texts workshop* (Vol. 3683, pp. 43–51). CEUR-WS. https://ceur-ws.org/Vol-3683/paper6.pdf

Grossner, K., Grunewald, S., & Mostern, R. (2023). Bringing places from the distant past to the present: A report on the world historical gazetteer. *International Journal on Digital Libraries*, *24*(3), 159–162. https://doi.org/10.1007/s00799-022-00341-2

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875–3889. https://doi.org/10.1098/rsta.2010.0149

Halterman, A. (2017). Mordecai: Full text geoparsing and event geocoding. *Journal of Open Source Software*, *2*(9), 91. https://doi.org/10.21105/joss.00091

Halterman, A. (2023). *Mordecai 3: A neural geoparser and event geocoder*. arXiv. https://doi.org/10.48550/arXiv.2303.13675

Hu, X., Kersten, J., Klan, F., & Farzana, S. M. (2024). Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. *International Journal of Geographical Information Science*, 1–28. https://doi.org/10.1080/13658816.2024.2405182

Karimzadeh, M., Pezanowski, S., MacEachren, A. M., & Wallgrün, J. O. (2019). GeoTxt: A scalable geoparsing system for unstructured text geolocation. *Transactions in GIS*, *23*(1), 118–136. https://doi.org/10.1111/tgis.12510

Kong, I., & Purves, R. S. (2025). Analyzing geographic bias of newspaper articles reporting global climate disasters. *Annals of the American Association of Geographers*, 1–19. https://doi.org/10.1080/24694452.2025.2564220

Purves, R. S., Clough, P., Jones, C. B., Hall, M. H., & Murdock, V. (2018). Geographic information retrieval: Progress and challenges in spatial search of text. *Foundations and Trends in Information Retrieval*, *12*(2-3), 164–318. https://doi.org/10.1561/1500000034

Teitler, B. E., Lieberman, M. D., Panozzo, D., Sankaranarayanan, J., Samet, H., & Sperling, J. (2008). NewsStand: A new view on news. *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. https://doi.org/10.1145/1463434.1463458

Yan, Y., & Lee, J. (2024). GeoReasoner: Reasoning on geospatially grounded context for natural language understanding. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4163–4167. https://doi.org/10.1145/3627673.3679934