A Large-Scale Dataset of Geographic Named Entities in PubMed

Xiaoliang Jiang, Vetle I. Torvik
School of Information Sciences, University of Illinois at Urbana-Champaign

1 Abstract

Although biomedical literature often includes geographic metadata, existing tools and data have limited performance and scalability. We present a dataset of extracted geographic named entities from 18.8 million abstracts in PubMed. The dataset contains 25,195,983 candidates extracted from 17,124,291 sentences across 7,956,499 articles. Each candidate is linked to an entity in GeoNames and MapAffil, and labeled with two probabilities, one based on an embedding language model, the other one based on contextual features, including metadata and linked data. Combining the two probabilities yields precision (93.2%) and recall (92.3%), which enable large-scale geographic analysis and science of science research of the biomedical literature.

The dataset is available from the Illinois Data Bank:
https://databank.illinois.edu/datasets/IDB-4044693?code=lrcRDrIA_zoV_4qBCwlzSpg6wojiVjn8hkpr5tAWZdw

2 Introduction

Geography has long shaped biomedical understanding of disease, public health, and knowledge exchange (Kirby et al., 2017). Large-scale bibliographic databases further support such studies (Lindberg & Humphreys, 2008), with PubMed a central resource due to its Geographical MeSH terms (Anders & Evans, 2010; Rosselli, 1999). These terms have facilitated research in geographic information retrieval (Purves et al., 2018), disaster response (Scalia et al., 2022), and epidemiological surveillance (Allen et al., 2017; Uthman & Uthman, 2008), as well as regional analyses in Nigeria (Adamson, 2025), Morocco (Badrane & Alaoui-el-Azher, 2003), Ivory Coast (Chatelin & Arvanitis, 2005), and Spain (Valderas et al., 2006). At the same time, Geographical MeSH is designed primarily at the article level and tends to emphasize broader regions, countries, or major cities (Tamames & de Lorenzo, 2010). While well-suited for many purposes, this scope can make it difficult to capture text-level geographic mentions (Weissenbacher et al., 2015). Additional challenges arise from biomedical vocabulary, abbreviations, author-defined terms, and complex entity structures (Jiang & Torvik, 2020; Sanati & Sanati, 2013), highlighting the need for scalable, automated geographic NER (Sanati & Sanati, 2013). Previous work has applied NER to historical texts (Won et al., 2018), social media (Middleton et al., 2018), web sources (Berragan et al., 2023), and general scientific publications (Acheson & Purves, 2021), demonstrating feasibility but leaving biomedical-specific issues less examined. Building on PubMed's Geographical MeSH while extending its coverage, our study introduces a dataset that integrates engineered features and predictive models to expose geographic information at the text level. Beyond training uses (weak supervision, NER, disambiguation), the dataset enables applications in public health surveillance, research provenance, and collaboration mapping, thereby supporting both NLP methods and equity-focused analyses of biomedical research.

3 Methodology

Our earlier work (Jiang et al., 2025) used a high-recall candidate harvest followed by a precision filter, but depended on multiple third-party NER tools, which limited throughput at PubMed scale. Here we adopt the same strategy—broad recall then refinement—but redesign both stages for efficiency and independence from external tools.

We begin with a rule-driven, fast-scan recall stage. Abstracts are sentence-segmented; within each sentence, surface cues (capitalization, span boundaries) generate candidate mentions, matched by longest-phrase and recursive

substring strategies against MapAffil and a restricted slice of GeoNames (A/P classes for countries/regions and settlements). Directional or administrative suffixes are stripped, and unsupported acronyms are discarded. The output is a candidate dataset (PMID, sentence text, order, dictionary hits) approximating "multi-tool union" recall at lower cost.

Features are grouped into three sets. (1) Textual: length, token counts, position, orthographic patterns, and lightweight context indicators (prepositions in/at/from; commas, parentheses, four-digit years; citation cues). (2) Embeddings: sentence encodings from BioBERT (Lee et al., 2019) and BioSimCSE (Kanakarajan et al., 2022), with BioSimCSE adopted (768-d). (3) Metadata/linked-data: MapAffil/GeoNames matches, GeoMeSH alignment, agreement with author and cited-author affiliations, cited GeoMeSH/affiliation, and global frequency statistics (log-transformed).

Preliminary supervision is provided by a conservative metadata-only rule layer ("silver standard"), based on dictionary confirmation, affiliations, GeoMeSH, frequency/consistency, and plausibility constraints. These high-precision positive labels only supervise Model A but are excluded from the final outputs.

Learning is cross-supervised. Model A: trained on silver labels using textual features and embeddings. Model B: trained on hard labels from Model A (e.g., 95/5 and 90/10 thresholds) using only metadata/linked-data features. This reduces label–feature entanglement. Model C: fuses Model A and B probabilities by log odds for calibrated scores.

The released dataset maps recognized candidates to standardized GeoNames identifiers. For each sentence, the candidate with the highest probability (probC) is selected as the primary result, linked to GeoNames (2025) and MapAffil (Torvik, 2015, 2018); hierarchically related mentions are merged into merged_candidate. Others are retained as other_potential_candidates. Each sentence thus yields one primary mapped candidate plus supplementary mentions and metadata.

4 Dataset Summary and Evaluation

Our dataset contains 25,195,983 candidates, and 17,124,291 primary candidates per sentence after hierarchical merging and deduplication from PubMed abstracts by 2018. Each record includes the PubMed identifier (PMID), the sentence with the place-name candidate, model probabilities (probA, probB9505, probC), and the affiliation-derived location field. In addition, the majority of records are linked to a GeoNames entity, over half include MapAffil information, and substantial portions also carry auxiliary signals such as geographic MeSH terms, cited geographic MeSH terms, alternative candidates, or merged candidates (see Table 1 for detailed counts).

Table 1. *Dataset Statistics and Coverage of Auxiliary Information*

Feature	Description	Count	(%)
PMID, sentence, Candidate, probA, probB9505, probC	Basic information and probabilities from models	17,124,291	100.00%
Og_mapaffil_res_best_geoname	Has the best-matched GeoNames result	17,117,565	99.96%
MapAffil_info_clean	Has MapAffil dictionary info	9,992,428	58.35%
geomesh_list	Has GeoMeSH terms	4,771,530	27.86%
cited_geomesh_list	Has cited GeoMeSH terms	9,393,275	54.85%
other_potential_candidates	Has other candidates with a lower probability	4,069,275	23.76%
merged_candidate	Has multiple consistent candidates	400,506	2.34%

Based on a manually checked gold-standard evaluation of 1,000 articles, our fast-scanning method achieved a 99.4% recall. After several rounds of model training, the final Model C reached a precision of 93.16%. This balance of

high recall and strong precision demonstrates the robustness and reliability of our approach for large-scale geographic entity recognition in biomedical literature. Table 2 shows a single record from our dataset with a description of each field.

Table 2. *Example Data Record with Candidate Extraction, Metadata, and Model Predictions*

Field	Value	Explanation
PMID	19028	PubMed identifier of the article
Sentence	"The unusually high relative frequency of cancer in the laryngeal region in males (18% of all histologically diagnosed cancers) and a sex ratio of unity for lung cancer in Northern Thailand were further explored in a hospital-based case-control study in Chiang Mai."	The sentence containing the place-name candidate
Candidate	Chiang Mai	Primary candidate with the highest probability
MapAffil_info_clean	[['18.790', '98.959', ['Chiang Mai, Thailand']]]	Geographic coordinates followed by a hierarchical place name mapped from the MapAffil dictionary
og_mapaffil_res_best_geoname	[1153670, 'ADM1', '18.75', '99.0', 'Chiang Mai Province', 'Thailand']	GeoNames ID, feature code, latitude, longitude, standardized name, and country
probA	0.991624	Probability from Model A (language-based)
probB9505	0.999993	Probability from Model B9505 (metadata-based)
probC	1.0	Final fused probability (Model C)
merged_candidate	Chiang Mai, Thailand	Two hierarchically related place names were detected and merged into one standardized candidate
other_potential_candidates	NaN	No additional candidates in this sentence
geomesh_list	['Thailand']	The article's MeSH terms include a Geographical MeSH entry for Thailand
unique_tempocity_level_list	[['Bukit Merah, Singapore'], ['Chiang Mai, Thailand'], ['Lyon, France'], ['Seattle, WA, USA'], ['nan']]	List of author affiliations in order; *nan* = missing value
cited_geomesh_list	['Thailand', 'Australia', 'United States', 'New Guinea', 'Berlin', 'India']	Geographical MeSH terms associated with cited references

Figure 1 illustrates the temporal coverage of article counts at three levels: all PubMed abstracts, articles in our dataset with at least one candidate place name, and those containing at least one high-probability geographic candidate (probC > 0.5), while the orange line tracks the fraction with geographic mentions. While the overall volume of publications has grown exponentially since the mid-20th century, the share of articles mentioning geographic entities has also steadily increased, particularly after the 1980s. Together, these figures indicate both the broader expansion of PubMed coverage and the rising salience of place-related dimensions in scientific research.

Figure 2 shows the temporal distribution of high-probability geographic candidates (probC > 0.5) across different granularities. The left y-axis indicates the relative share of four candidate types (country, state, city, region), while the right y-axis (black line, log10 scale) tracks the total number of high-probability candidates per year.

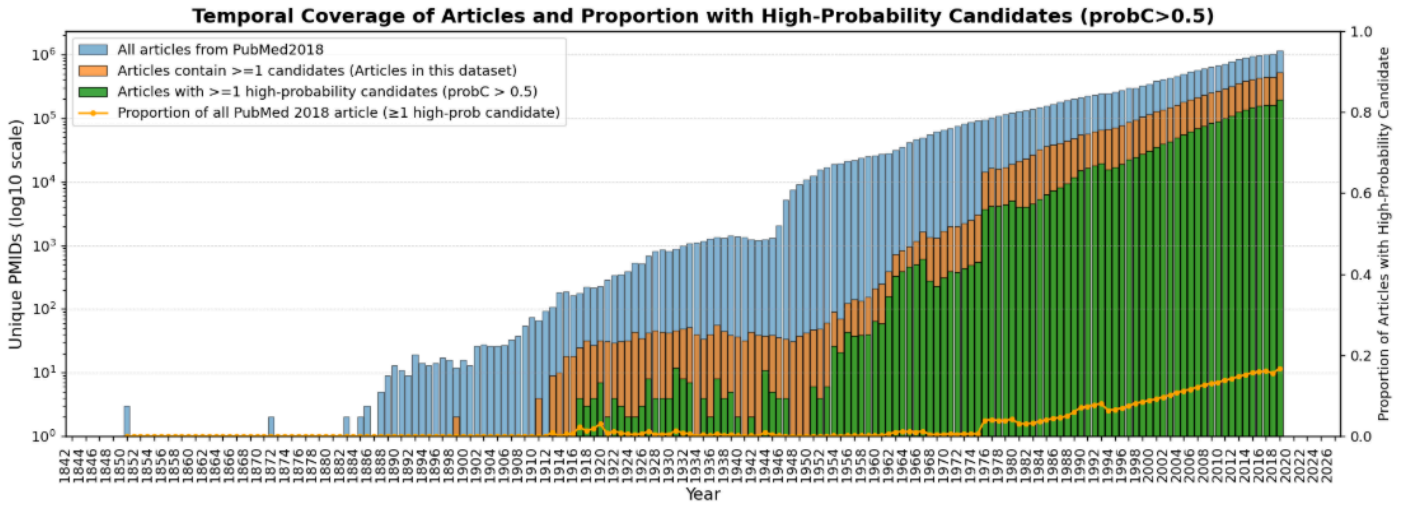


Figure 1. *Temporal Coverage of Articles and Proportion with High-Probability Candidates (probC>0.5)*

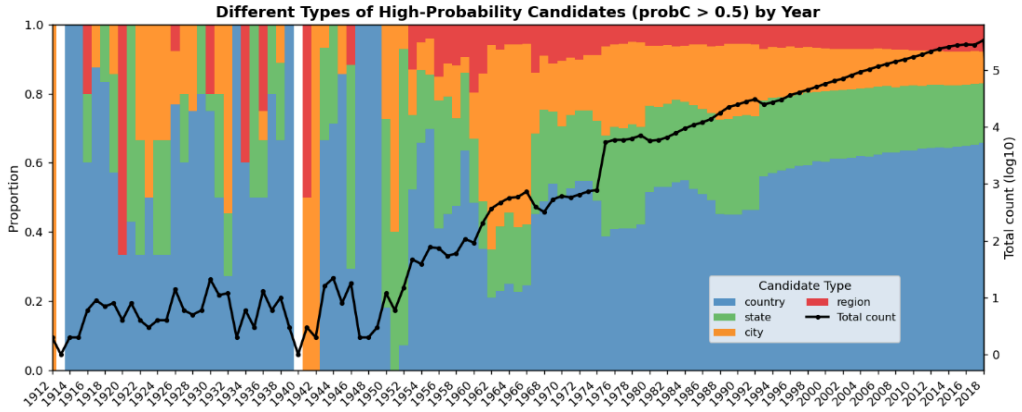


Figure 2. *Temporal Trends and Overall Composition of High-Probability Candidates (probC > 0.5)*

Before 1960, sparse data led to small volumes and unstable proportions across granularities. With the post-1960 surge in publications, distributions stabilized: countries dominated, states and cities held consistent shares, and regions remained minor. Across models, countries account for most high-confidence positives (61.5%), followed by states (19.9%), cities (11.3%), and regions (7.2%). These figures together highlight both the temporal dynamics and the overall composition of geographic candidates.

Figure 3 illustrates how textual and metadata-based models provide complementary yet imperfect signals for geographic named entity recognition in biomedical literature. Model A, trained on textual features, yields relatively smooth probability distributions that capture sentence-level linguistic cues, but it may produce errors in cases involving abbreviations, author-specific terminology, or citation expressions. Model B, derived from metadata-based labels, exhibits discontinuities resulting from binary indicators such as affiliations and GeoMeSH terms. Although largely article-level in scope, its incorporation of candidate-level consistency checks enables it to elevate mentions aligned with metadata, occasionally introducing spurious positives such as “CT” (Connecticut or Computed Tomography). Model C integrates A

and B through a log-odds formulation, concentrating on high-confidence predictions and reinforcing cases where textual and metadata evidence converge, while attenuating uncertainty when they diverge. Collectively, these models highlight the challenge of reconciling noisy textual signals with metadata that is informative yet prone to over-interpretation, and demonstrate that combining the two approaches can enhance precision by filtering out borderline cases.

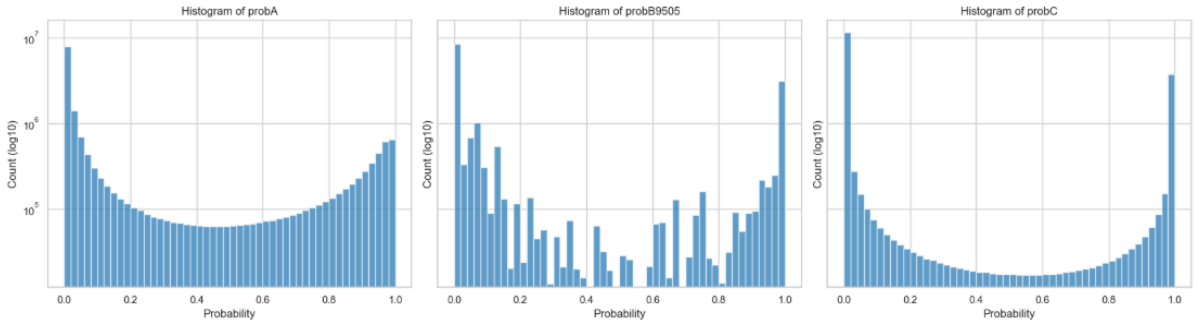


Figure 3. *Probability Distributions of High-Confidence Candidates Across Models A, B9505, and C*

5 Discussion

Our analysis reveals clear temporal and categorical trends in the distribution of geographic information in PubMed, showing that such information, once sparse and inconsistent, has become a stable and increasingly salient dimension of biomedical literature. Nearly twenty percent of articles contain at least one place name in abstracts, and the dataset enables diverse applications, from global health monitoring and analyses of regional disparities to studies of research collaborations through the linkage of place mentions and author affiliations. Figure 4, for example, illustrates how the first author's country (top 20 as indicated in the legend) relates to geographic mentions (shown as positions on the map with candidate counts represented by size), with the United States not only leading in publication volume but also spanning regions such as Africa, the Middle East, Southeast Asia, and Eastern Europe—patterns that echo ongoing debates on research equity (Else, 2025). By providing systematic text-level identification of geographic information, our dataset offers a benchmark for machine learning as well as evidence to investigate the global distribution and equity of biomedical research.

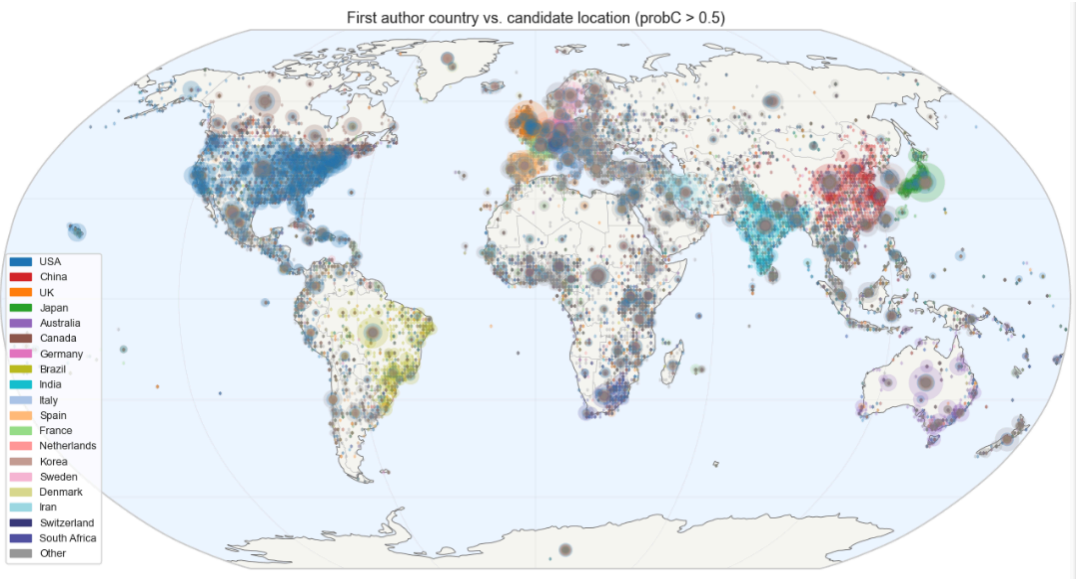


Figure 4. *Global Distribution of Candidate Places (probC>0.5) by First Author Country (Colored)*

Acknowledgement

Research reported in this publication was supported by the US National Institutes of Health (Award Number P01AG039347). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Acheson, E., & Purves, R. S. (2021). Extracting and modeling geographic information from scientific articles. *PLOS ONE*, *16*(1), e0244918. https://doi.org/10.1371/journal.pone.0244918

Adamson, I. (1992). Access and retrieval of information as coordinates of scientific development and achievement in Nigeria. *Scientometrics*, *23*(1), 191–199. https://doi.org/10.1007/BF02020922

Allen, T., Murray, K. A., Zambrana-Torrelio, C., Morse, S. S., Rondinini, C., Di Marco, M., Breit, N., Olival, K. J., & Daszak, P. (2017). Global hotspots and correlates of emerging zoonotic diseases. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00923-8

Anders, M. E., & Evans, D. P. (2010). Comparison of PubMed and Google Scholar Literature Searches. *Respiratory Care*, *55*(5), 578–583. https://doi.org/10.4187/respcare.10550578

Badrane, H., & Alaoui-el-Azher, M. (2003). Biomedical research in developing countries: The case of Morocco in the 1990s: La Tunisie medicale. *La Tunisie Medicale*, *81*(6), 377–382.

Berragan, C., Singleton, A., Calafiore, A., & Morley, J. (2023). Transformer based named entity recognition for place name extraction from unstructured text. *International Journal of Geographical Information Science*, *37*(4), 747–766. https://doi.org/10.1080/13658816.2022.2133125

Chatelin, Y., & Arvanitis, R. (2005). Representing scientific activity by structural indicators: The case of Cote d'Ivoire 1884–1968. *Scientometrics*, *23*(1), 235–247. https://doi.org/10.1007/bf02020925

Else, H. (2025). Top universities warned against unfair research partnerships on their doorstep. *Nature*. https://doi.org/10.1038/d41586-025-00159-3

GeoNames. (2025). Retrieved Sep 24, 2025, from https://www.geonames.org/

Jiang, X., Bosch, N., & Torvik, V. I. (2025). Training a Geographic Entity Recognizer on Biomedical Abstracts with the Aid of Embeddings, Metadata, and Linked Data. *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, 1–5. https://doi.org/10.1145/3677389.3702515

Jiang, X., & Torvik, V. I. (2020). On the Ambiguity and Relevance of Place Names in Scientific Text. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 401–404. https://doi.org/10.1145/3383583.3398618

Kanakarajan, K. raj, Kundumani, B., Abraham, A., & Sankarasubbu, M. (2022). BioSimCSE: BioMedical Sentence Embeddings using Contrastive learning. *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, 81–86. https://aclanthology.org/2022.louhi-1.10

Kirby, R. S., Delmelle, E., & Eberth, J. M. (2017). Advances in spatial epidemiology and geographic information systems. *Annals of Epidemiology*, *27*(1), 1–9. https://doi.org/10.1016/j.annepidem.2016.12.001

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, btz682. https://doi.org/10.1093/bioinformatics/btz682

Lindberg, D. a. B., & Humphreys, B. L. (2008). Rising Expectations: Access to Biomedical Information. *Yearbook of Medical Informatics*, *17*(1), 165–172. https://doi.org/10.1055/s-0038-1638596

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location Extraction from Social Media: Geoparsing, Location Disambiguation, and Geotagging. *ACM Transactions on Information Systems*, *36*(4), 40:1-40:27. https://doi.org/10.1145/3202662

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

Rosselli, D. (1999). Geography of biomedical publications. *The Lancet*, *354*(9177), 517. https://doi.org/10.1016/S0140-6736(05)75555-0

Sanati, N. A., & Sanati, M. (2013). Growing interest in use of geographic information systems in health and healthcare research: A review of PubMed from 2003 to 2011. *JRSM Short Reports*, *4*(6), 2042533313478810. https://doi.org/10.1177/2042533313478810

Scalia, G., Francalanci, C., & Pernici, B. (2022). CIME: Context-aware geolocation of emergency-related posts. *GeoInformatica*, *26*(1), 125–157. https://doi.org/10.1007/s10707-021-00446-x

Tamames, J., & de Lorenzo, V. (2010). EnvMine: A text-mining system for the automatic extraction of contextual information. *BMC Bioinformatics*, *11*(1), 294. https://doi.org/10.1186/1471-2105-11-294

Torvik, V. I. (2015). MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide. *D-Lib Magazine : The Magazine of the Digital Library Forum*, *21*(11–12), 10.1045/november2015-torvik.

Torvik, V. (2021). *MapAffil 2018 dataset—PubMed author affiliations mapped to cities and their geocodes worldwide with extracted disciplines, inferred GRIDs, and assigned ORCIDs*. https://doi.org/10.13012/B2IDB-2556310_V1

Uthman, O. A., & Uthman, M. B. (2008). Geography of Africa biomedical publications: An analysis of 1996–2005 PubMed papers. *African Journal of Food, Agriculture, Nutrition and Development*, *8*(2), Article 2. https://doi.org/10.4314/ajfand.v8i2.19192

Valderas, J. M., Mendivil, J., Parada, A., Losada-Yáñez, M., & Alonso, J. (2006). Development of a Geographic Filter for PubMed to Identify Studies Performed in Spain. *Revista Española de Cardiología (English Edition)*, *59*(12), 1244–1251. https://doi.org/10.1016/S1885-5857(07)60080-2

Weissenbacher, D., Tahsin, T., Beard, R., Figaro, M., Rivera, R., Scotch, M., & Gonzalez, G. (2015). Knowledge-driven geospatial location resolution for phylogeographic models of virus migration. *Bioinformatics*, *31*(12), i348–i356. https://doi.org/10.1093/bioinformatics/btv259

Won, M., Murrieta-Flores, P., & Martins, B. (2018). Ensemble Named Entity Recognition (NER): Evaluating NER Tools in the Identification of Place Names in Historical Corpora. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018.00002