Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers

Jimin Wang | Yingjie Hu

GeoAI Lab, Department of Geography, University at Buffalo, Buffalo, New York

Correspondence
Yingjie Hu, GeoAI Lab, Department of Geography, University at Buffalo, Buffalo, NY 14260.
Email: yhu42@buffalo.edu

Abstract
A rich amount of geographic information exists in unstructured texts, such as web pages, social media posts, housing advertisements, and historical archives. Geoparsers are useful tools that extract structured geographic information from unstructured texts, thereby enabling spatial analysis on textual data. While a number of geoparsers have been developed, they have been tested on different data sets using different metrics. Consequently, it is difficult to compare existing geoparsers or to compare a new geoparser with existing ones. In recent years, researchers have created open and annotated corpora for testing geoparsers. While these corpora are extremely valuable, much effort is still needed for a researcher to prepare these data sets and deploy geoparsers for comparative experiments. This article presents EUPEG: an Extensible and Unified Platform for Evaluating Geoparsers. EUPEG is an open source and web-based benchmarking platform which hosts the majority of open corpora, geoparsers, and performance metrics reported in the literature. It enables direct comparison of the geoparsers hosted, and a new geoparser can be connected to EUPEG and compared with other geoparsers. The main objective of EUPEG is to reduce the time and effort that researchers have to spend in preparing data sets and baselines, thereby increasing the efficiency and effectiveness of comparative experiments.
INTRODUCTION

Many studies and applications nowadays need an integration of spatial and textual analysis. In disaster response, it is often necessary to recognize, geolocate, and analyze the place names mentioned in short text messages or social media posts in order to understand who needs help and where (Gelernter & Balaji, 2013; Lan, Adelfio, & Samet, 2014; MacEachren et al., 2011; Pezanoski, MacEachren, Savelyev, & Robinson, 2018). Studying place relations and interactions in virtual or cognitive spaces usually involves extracting place names from texts, such as Wikipedia pages or news articles, and analyzing their co-occurrences with spatial distances (Geiß, Spitz, Strötgen, & Gertz, 2015; Hecht & Moxley, 2009; Hu, Ye, & Shaw, 2017; Liu, Wang, Kang, Gao, & Lu, 2014; Salvini & Fabrikant, 2016). To develop place-based GIS, researchers may need to examine human experiences encoded in texts, such as in travel blogs, and how these human experiences are related to spatial locations (Adams & McKenzie, 2013; Ballatore & Adams, 2015; Gao et al., 2017). In addition, there exists a vast amount of geographic knowledge in web pages (Jones, Purves, Clough, & Joho, 2008), housing advertisements (McKenzie, Liu, Hu, & Lee, 2018), business documents (Faulconbridge, Hall, & Beaverstock, 2008), historical archives (Grossner, Janowicz, & Keßler, 2016), and other types of texts.

Geoparsing is a critical process for extracting spatial information from textual data. It is recognized as an important research topic in the broader field of geographic information retrieval (GIR; Jones & Purves, 2008; Purves, Clough, Jones, Hall, & Murdock, 2018). A geoparsing system which takes unstructured textual data as the input and outputs a set of recognized place names and their spatial footprints is called a geoparser (Freire, Borbinha, Calado, & Martins, 2011; Leidner, 2008).

While a number of geoparsers have already been developed, it is difficult to directly compare their performances. Examples of geoparsers developed include MetaCarta (Frank, Rauch, & Donoghue, 2006), GeoTxt (Karimzadeh et al., 2013), the Edinburgh Geoparser (Alex, Byrne, Grover, & Tobin, 2015), TopoCluster (DeLozier, Baldridge, & London, 2015), and CamCoder (Gritta, Pilehvar, & Collier, 2018b). Two factors make direct comparison difficult. First, many geoparsers have been tested on project-specific data sets that are not shared publicly. Besides the additional effort of making data ready for sharing, there also exist policy restrictions (e.g. Twitter forbids sharing the content of tweets) and privacy concerns that prevent researchers from sharing their data publicly. As a result, geoparsers cannot be fairly compared since the same geoparser can exhibit very different performance depending on the data sets used for testing (Ju et al., 2016; Leidner, 2006). Second, different performance metrics can be used for evaluating geoparsers. Some researchers use the metrics of precision, recall, and F-score adopted from the field of information retrieval, while others use metrics based on spatial distances, such as mean or median error distance. For these reasons, we cannot compare geoparsers by juxtaposing the performance numbers reported in their papers.

To effectively compare existing geoparsers or to compare a new geoparser with existing baselines, one would ideally find and deploy the geoparsers in the literature and use the same data sets and metrics to test their performance. However, such a process would be time-consuming and labor-intensive. First, one would need to find openly shared and annotated data sets. While the community has already made great efforts to share data sets, such as the Local-Global Lexicon (LGL) corpus (Lieberman, Samet, & Sankaranarayanan, 2010), WikToR (Gritta, Pilehvar, Limsoopatham, & Collier, 2018), and GeoCorpora (Wallgrün, Karimzadeh, MacEachren, & Pezanoski, 2018), researchers still need to spend much time preparing these data sets for experiments. For example, GeoCorpora is a valuable data set containing human-annotated tweets. Due to Twitter’s policy restrictions, GeoCorpora contains only the IDs of tweets rather than their full content. To use GeoCorpora, one needs to apply for a developer account to use Twitter’s application programming interface (API) and write a program to rehydrate the data set. When data sets are more readily available they are often in different formats and need to be harmonized into the same format before an experiment. Second, it takes a considerable amount of time to find, deploy, and rerun existing geoparsers. This process can take even longer when no direct download link is provided for a geoparser or when there is a lack of deployment instructions. Third, a set of performance metrics must be implemented to compare geoparsers. Although implementing these metrics may not be difficult, one needs to harmonize the
heterogeneous output formats of different geoparsers and compare their outputs with ground-truth annotation. In sum, conducting an effective comparative experiment of geoparsers costs a lot of time and human resources. While those costs are probably fine for a single research group, the community as a whole can waste a lot of time if every individual research group has to prepare data sets, geoparsers, and metrics in order to run an experiment.

This article presents EUPEG: an Extensible and Unified Platform for Evaluating Geoparsers. EUPEG is designed as an open source and web-based platform. It hosts the majority of the open corpora, geoparsers, and performance metrics reported in the literature. One can directly compare the geoparsers hosted on the same data sets using the same metrics, or connect a new geoparser to EUPEG and compare it with the existing ones. The value of EUPEG can be seen from the perspectives of both geoparser users and researchers. For a user who would like to find a suitable geoparser to process a corpus, EUPEG offers a comprehensive view on the advantages and limitations of different geoparsers (e.g. some may have higher precision while some others may have higher recall) and their performance on different types of corpora (e.g. short messages or long articles). Researchers who would like to develop a new geoparser can focus on inventing new methods rather than preparing data sets and baselines. In addition, EUPEG automatically archives the results and configurations of experiments, such as the date and time of an experiment, the data sets selected, geoparsers used, and performance metrics. Researchers can share experiment results with their colleagues or even the general public more easily via an experiment ID.

The contributions of this article are as follows:

- We propose and develop a benchmarking platform, EUPEG, for effective and efficient comparison of geoparsers. EUPEG currently hosts eight annotated geographic corpora, nine geoparsers, and eight performance metrics. New geoparsers and data sets can also be connected to it. Experiment results and configurations are recorded and can be shared via experiment IDs. A demo of EUPEG can be accessed at https://geoai.geog.buffalo.edu/EUPEG.
- We provide a systematic review of the geoparsing resources hosted on EUPEG. The corpora are in four different text genres, ranging from news articles to social media posts; the geoparsers are developed using different methods, such as heuristics and machine learning; and the performance metrics are from information retrieval or based on spatial distances. EUPEG serves as a one-stop platform that unifies the heterogeneous data sets, geoparsers, and performance metrics.
- We share the EUPEG source code on GitHub, along with the hosted resources under permitted licenses (e.g. GNU General Public License). The code repository can be accessed at https://github.com/geoai-lab/EUPEG. The source code shared enables researchers to run EUPEG on a local computer, or to add more data sets and geoparsers at the source-code level. One can also extend EUPEG with new features, such as new performance metrics suitable for a project.

While EUPEG is designed for geoparsing, the idea of developing benchmarking platforms can be extended to other research topics in geography, such as land use and land cover classification, where different solutions are developed for addressing the same problem. This article is a major extension of our previous short paper (Hu, 2018). The remainder of this article is organized as follows. Section 2 reviews related work on geoparsing, corpus building, and benchmarking platforms. Section 3 presents the design details of EUPEG, including the overall architecture and the hosted data sets, geoparsers, and performance metrics. Section 4 demonstrates the EUPEG implementation and provides an analytical evaluation on the approximate time that can be saved by EUPEG for comparative experiments. Section 5 provides a summary and discusses future directions.

2 | RELATED WORK

In this section we provide a review on related studies. We start by introducing the background to geoparsing and the major geoparsers developed so far, and go on to discuss the efforts made by the community to create and
share open and annotated corpora. We then discuss the recent movement toward the development of benchmarking platforms for effective and efficient comparisons of different solutions to the same problems.

2.1 Geoparsing and geoparsers

Geoparsing is a research topic often studied in GIR (Jones & Purves, 2008; Purves et al., 2018). The goal of geoparsing is to recognize place names mentioned in texts and resolve them to the corresponding place instances and location coordinates (Barbaresi, 2017; Freire et al., 2011; Gritta et al., 2018). Geoparsing is typically performed in two consecutive steps: toponym recognition and toponym resolution. The first step recognizes place names from texts without identifying the particular place instance referred to by a name, while the second step aims to resolve any ambiguity of the place name and locate it to the right spatial footprint. Figure 1 illustrates the input and output of geoparsing and its two steps. Many methods have been proposed for these two steps. For toponym recognition, early research used gazetteer-based entry matching and grammatical rules (Purves et al., 2018; Woodruff & Plaunt, 1994), while more recent approaches have employed machine learning and natural language processing techniques. In particular, the Stanford Named Entity Recognition (NER) tool has been used in many studies for toponym recognition (DeLozier et al., 2015; Gelernter & Mushegian, 2011; Karimzadeh et al., 2013; Lieberman et al., 2010). For toponym resolution, various heuristics have been developed to resolve place name ambiguity (Amitay, Har’El, Sivan, & Soffer, 2004; Leidner, 2008). A simple method is to resolve a place name to the place instance that has the highest population or the largest total geographic area (Ladra, Luaces, Pedreira, & Seco, 2008; Li, Sripri, Niu, & Li, 2002). Machine learning models have also been developed for toponym resolution by exploiting various features, such as toponym co-occurrences (Overell & Rüger, 2008), words in the local context (Speriosu & Baldridge, 2013), distances among toponyms (Santos, Anastácio, & Martins, 2015), topics of the local context (Ju et al., 2016), and a combination of multiple features (Gritta et al., 2018b; Nesi, Pantaleo, & Tenti, 2016).

A number of geoparsers have been developed which can function as end-to-end systems for completing both steps. GeoTxt, developed by Karimzadeh et al. (2013), is a web-based geoparser that leverages Stanford NER and two other NER tools for toponym recognition and uses GeoNames and a set of heuristic rules for toponym resolution. TopoCluster, by DeLozier et al. (2015), can perform geoparsing without using a gazetteer. It uses Stanford NER to recognize toponyms from texts and then resolves toponyms based on the geographic profiles of words in the surrounding context. The Cartographic Location and Vicinity Indexer (CLAVIN) is an open source geoparser that employs Apache OpenNLP for toponym recognition and utilizes a gazetteer and fuzzy search for toponym resolution. The Edinburgh Geoparser is a geoparsing system developed by the Language Technology Group at Edinburgh University (Alex et al., 2015). It uses their in-house software tool, called LT-TTT2, for toponym recognition, and the toponym resolution step is based on a gazetteer such as GeoNames. CamCoder is a deep-learning-based geoparser developed by Gritta et al. (2018b), which integrates convolutional neural networks, word embeddings, and the geographic vector representations of place names. There also exist commercial geoparsers, such as Geoparser.io (https://geoparser.io), which often charge a fee. Some commercial geoparsers, such

FIGURE 1 The input and output of geoparsing and its two main steps
as Yahoo! PlaceSpotter (https://developer.yahoo.com/boss/geo/docs/PM_KeyConcepts.html), provide relatively permissive rate limitations for free requests (e.g. 2,000 calls per hour).

2.2 | Sharing open and annotated corpora

While many geoparsers exist, it is difficult to directly compare them due to a lack of open and annotated corpora. In recent years, researchers have made great efforts to address this issue. Lieberman et al. (2010) shared a human-annotated data set called LGL containing 588 news articles published by local newspapers from highly ambiguous places. Hu, Janowicz, and Prasad (2014) contributed an automatically annotated corpus containing short sentences retrieved from the home pages of cities with ambiguous names such as Washington. Ju et al. (2016) shared a corpus of short sentences from various web pages, which were automatically collected and annotated using a script based on the Microsoft Bing Search API. Gritta et al. (2018) contributed WikToR, which is a corpus of Wikipedia articles with ambiguous names, such as Lima, Peru and Lima, Oklahoma, automatically annotated by a Python script. Wallgrün et al. (2018) contributed GeoCorpora, which is a data set of tweets manually annotated using a hybrid approach with both users on Amazon's Mechanical Turk and researchers in the domain of geography. Gritta, Pilehvar, and Collier (2018a) and Gritta et al. (2018b) shared two human annotated corpora, GeoVirus and GeoWebNews, which contain 229 and 200 news articles, respectively. TR-News is another news article corpus which contains 118 articles manually annotated by Kamalloo and Rafiei (2018). In addition to contemporary corpora, some historical data sets were also made available, such as War of the Rebellion (WOTR) by DeLozier, Wing, Baldridge, and Nesbit (2016). Leidner (2006) developed the TR-CoNLL corpus which contains 946 annotated news articles from Reuters; however, to the best of our knowledge, it is not publicly available. The ACE 2005 English SpatialML is an annotated news corpus shared on the Linguistic Data Consortium (Mani, Hitzeman, Richer, & Harris, 2008), but it charges a fee ($1,000) for non-members. While these annotated corpora greatly facilitate the development and testing of geoparsers, considerable amounts of time and effort are required to find, download, and prepare these corpora.

2.3 | Benchmarking platforms

The importance and necessity of evaluating geoparsers in a systematic manner have already been recognized by the research community (Gritta et al., 2018; Melo & Martins, 2017; Monteiro, Davis, & Fonseca, 2016; Richter, Geiß, Spitz, & Gertz, 2017; Wallgrün et al., 2018). Melo and Martins (2017) argued that the fact that one geoparser performed worse than another geoparser on one particular data set did not mean that it would perform worse if a different data set were used. Gritta et al. (2018) compared the performance of five geoparsers on two corpora using a set of standard performance measures, such as precision, recall, F-score, and median error distance. In their more recent work, the authors further proposed a pragmatic guide to geoparsing evaluation (Gritta et al., 2018a). In addition to publications, Gritta et al. (2018a, 2018) also released their source code and the annotated corpora which greatly facilitated the reproduction of their experiments. EUPEG is built on the foundational work of Gritta et al. (2018a, 2018), but extends their work in three respects. First, EUPEG provides a benchmarking platform which offers data sets and baseline geoparsers that are ready for use. While Gritta et al. (2018a, 2018) shared the source code for comparing five geoparsers, a lot of effort is still needed to understand, deploy, and run these geoparsers. Some geoparsers do not function on certain operating systems (e.g. the Edinburgh Geoparser is not supported on Windows) or require extra database configurations (e.g. TopoCluster requires PostgreSQL and PostGIS), which can add additional requirements for their deployment. EUPEG directly hosts these geoparsers, along with annotated corpora and performance metrics. Researchers can directly run experiments on EUPEG, and can connect their own geoparsers and data sets to the platform. Second, EUPEG extends the corpora and geoparsers from Gritta et al. (2018a, 2018). We provide eight annotated corpora in four different text genres: news articles, Wikipedia articles, social media posts, and web pages. We provide nine geoparsing methods which
include not only specialized geoparsers (e.g. GeoTxt and CLAVIN) but also a number of geoparsing systems extended from general NER tools, such as DBpedia Spotlight, Stanford NER, and spaCy NER. Third, EUPEG offers the capability of archiving research experiments. Each experiment is assigned a unique ID that allows researchers to share first-hand research outcomes and to search the results of previous experiments.

The demand for benchmarking platforms is also witnessed in other research fields beyond geography. Cornolti, Ferragina, and Ciaramita (2013) developed a framework for systematically evaluating a number of named entity annotators, such as AIDA, Illinois Wikifier, and DBpedia Spotlight, on the same data sets. Building on the work of Cornolti et al. (2013), Usbeck et al. (2015) developed GERBIL, which is a platform for agile, fine-grained, and uniform evaluations of named entity annotation tools. The practice of comparing different methods on the same data sets has also been seen at computer science conferences, such as the Message Understanding Conference (MUC; Sundheim, 1993), the Conference on Computational Natural Language Learning (CoNLL; Tjong Kim Sang & De Meulder, 2003), and the Making Sense of Microposts (MSM) workshop series (Cano et al., 2014). Such practice is especially effective when multiple solutions exist for the same research problem, and can reveal the advantages and limitations of different solutions. Sharing data sets for comparing methods can fuel advancement in a particular research area as well. For example, the availability of the ImageNet data set was a critical boost to the remarkable development of deep learning in computer vision (Deng et al., 2009). To the best of our knowledge, EUPEG is the first benchmarking platform for the research problem of geoparsing.

3 | EUPEG

3.1 | Overall architecture

EUPEG is designed as a web-based and open source benchmarking platform. It provides two main functions:

- It enables effective and efficient comparison of different geoparsers on the same data sets using the same performance metrics.
- It facilitates the sharing of experiments by archiving evaluation results and configurations and supporting the search of previous experiments.

The overall architecture of EUPEG is shown in Figure 2. Two major modules are designed. The Experiment Module hosts the majority of openly available resources, including annotated corpora, existing geoparsers, and performance metrics. The Archiving and Search Module records experiment results and supports the search of previous experiments. Geoparser researchers who would like to develop new geoparsers can connect a new geoparser to EUPEG and compare it with others on the data sets hosted using the same performance metrics. Researchers can also upload one or more customized data sets and compare the new geoparser with existing baselines on these customized data sets. Users who would like to find a suitable geoparser for processing a corpus can compare existing geoparsers directly on EUPEG to see their advantages and limitations. The experiment configurations (e.g. the data sets, geoparsers, and metrics used) and results are recorded in a database in the Archiving and Search Module. One can search previous experiments based on their experiment IDs automatically generated by EUPEG. In the following, we present details of the data sets, geoparsers, and metrics.

3.2 | Data sets

EUPEG hosts the majority of annotated geographic corpora reported in the literature. Two criteria have been used for selecting these data sets. First, they have to be formally described by existing papers. Second, the data sets should be openly available without a fee (e.g. a data set shared on GitHub under the MIT license). In addition, we focus on geographic corpora, namely those with toponyms and spatial footprints annotated. There exist general
NER corpora whose annotations contain other types of entities (e.g., persons and concepts) and do not provide spatial footprints. Those general NER corpora are not included.

The corpora hosted on EUPEG are in four different text genres: news articles, Wikipedia articles, social media posts, and web pages. Having multiple genres rather than a single text type can provide a more comprehensive evaluation on the performance of a geoparser (Leidner, 2006). These data sets are described as follows.

**News articles.** Four news article corpora are hosted on EUPEG:

- **LGL.** LGL, developed by Lieberman et al. (2010), contains 588 human-annotated news articles published by 78 local newspapers from highly ambiguous places, such as *Paris News* (Texas), *Paris Post-Intelligencer* (Tennessee), and *Paris Beacon-News* (Illinois).
- **GeoVirus.** This is a human-annotated data set shared by Gritta et al. (2018b). It contains 229 news articles from WikiNews during August and September 2017. These news articles cover global disease outbreaks and epidemics, and were collected using keywords such as *Ebola* and *AIDS*.
- **TR-News.** This data set, contributed by Kamalloo and Rafiei (2018), contains 118 human-annotated news articles from various global and local news sources. The authors deliberately included less dominant place name instances, such as *Edmonton, England* and *Edmonton, Australia*, while also keeping articles from general global news sources, such as BBC and Reuters.
- **GeoWebNews.** This data set was shared by Gritta et al. (2018a) and comprises 200 human-annotated news articles from 200 globally distributed news sites collected during April 1–8, 2018. The authors randomly selected one article from each domain until they reached 200 news articles.

**Wikipedia articles.** One Wikipedia article corpus, *WikToR*, is hosted on EUPEG. This corpus was provided by Gritta et al. (2018). It was automatically generated by a script using Wikipedia articles about places. It contains 5,000 articles.
with ambiguous names, such as Lima, Peru, Lima, Ohio, and Lima, Oklahoma. One limitation of WikToR is that it does not annotate all toponyms in the texts but only those that are the description target of a Wikipedia article. As a result, some performance metrics, such as precision, recall, and F-score, cannot be used to quantify the performance of geoparsers based on WikToR.

Social media posts. EUPEG hosts one social media data set, GeoCorpora, contributed by Wallgrün et al. (2018). This is a tweet corpus that contains 1,639 human-annotated tweet posts. These posts were first annotated using a crowdsourcing approach by workers on Amazon's Mechanical Turk and then further annotated (with disagreements resolved) by researchers in geography. It is worth noting that the original paper reported 2,122 tweets with toponyms annotated. Due to Twitter's data sharing restrictions, Wallgrün et al. (2018) could not share the full content of tweets but only tweet IDs. When rehydrating tweets, we were able to recover only 1,639 of them, since some tweets were deleted by their authors. The number of tweets that can be recovered will only decrease as time goes by. Thus, we believe that another value of EUPEG is its capability of preserving valuable contributions from previous research by sidestepping some policy restrictions (in this case, EUPEG does not provide any direct download of the tweets).

Web pages. Two web page corpora are hosted on EUPEG:

- **Hu2014.** This is a small corpus, contributed by Hu et al. (2014), which was automatically constructed by a script. The authors focused on two highly ambiguous US place names, Washington and Greenville, and retrieved textual descriptions from the websites of related cities. The texts in these web pages were then divided into shorter sentences. Overall, this data set contains 134 entries. Not all toponyms in the sentences are annotated, and therefore precision, recall, and F-score cannot be used to evaluate the geoparsing results based on this data set (similarly to WikToR).
- **Ju2016.** This is another automatically constructed corpus. It was contributed by Ju et al. (2016), who made use of a list of highly ambiguous US place names on Wikipedia, and then used the Microsoft Bing Search API to retrieve sentences from various web pages (Wikipedia articles were removed from these web pages) that contain the place names searched. The corpus contains 5,441 entries. Similar to WikToR and Hu2014, this data set does not annotate all toponyms and cannot use the performance metrics of precision, recall, and F-score.

In total, EUPEG hosts eight geographic corpora in four different text genres. Table 1 summarizes the attributes of these data sets.

There also exist open and annotated historical corpora. For example, WOTR is a U.S. civil war corpus with toponyms focusing on the southern USA (e.g. DeLozier et al., 2016). However, geoparsing such corpora requires special configurations such as adding historical gazetteers and processing older languages (e.g. the texts of WOTR)

| Table 1 | A summary of the open and annotated corpora hosted on EUPEG |
|---------|-------------------------------------------------------------|
| Data set | Genre | Text date | Entry count | Average words per entry | Average toponyms per entity |
| LGL     | News  | 03/2009  | 588        | 315                  | 8.0                     |
| GeoVirus| News  | 08–09/2017| 229        | 276                  | 9.4                     |
| TR-News | News  | 2009–2017| 118        | 324                  | 10.8                    |
| GeoWebNews| News | 04/2018  | 200        | 404                  | 12.6                    |
| WikToR  | Wikipedia | 03/2016 | 5,000     | 213                  | 6.3                     |
| GeoCorpora | Social media | 2014–2015 | 1,639      | 19                   | 2.1                     |
| Hu2014  | Web pages | 08/2014 | 134        | 27                   | 1.3                     |
| Ju2016  | Web pages | 11/2016 | 5,441     | 21                   | 1.2                     |
are from the 1860s). The current version of EUPEG aims to compare geoparsers based on their default configurations and does not include historical corpora.

It is worth noting that the definition of “toponym” can vary across different corpora. It seems that different researchers often have their own opinions on what should be considered as toponyms and what should not. This definition difference affects the ground-truth toponym annotation in a corpus. For example, LGL considers demonyms (e.g. Canadian) as toponyms and annotates them to point coordinates (e.g. the center of Canada), whereas GeoVirus does not annotate building names, point-of-interest names, street names, and river names. In a recent work, Gritta et al. (2018a) provided a pragmatic taxonomy of toponyms which further divided toponyms into literal and associative toponyms with 13 subcategories. In this work, we do not attempt to define “toponym” from a single perspective, but allow data sets with different toponym definitions to coexist. Such diversity allows users and researchers to see the different performance of a geoparser across corpora. They can then choose a geoparser that performs best on a corpus that has a toponym definition similar to their own. Table 2 summarizes the different annotations of toponyms contained in each of the corpora hosted on EUPEG.

As can be seen, all data sets include administrative units in their annotations but have different coverage on other types of entities. Some of these differences come from the corpus building process (e.g. WikToR, Hu2014, and Ju2016 were automatically constructed based on the names of cities and towns only), while some others originate from the different views of researchers on the definition of toponym. It seems that domain knowledge plays a major role in the annotation of toponyms. For example, GeoCorpora is a data set contributed by geographers, and its annotations contain only the names that can be pinned down to a certain location on the surface of the Earth. By contrast, GeoWebNews, TR-News, and LGL are contributed by linguists and computer scientists who tend to annotate any terms that may have a geographic meaning (e.g. “Canadian” and “Spanish sausage”).

### 3.3 Geoparsers

We select geoparsers for EUPEG using the following criteria. First, they should function as end-to-end systems, that is, they should be able to take textual documents as the input and output spatial coordinates. Second, for academic geoparsers, the accompanying papers should be published after 2010 and they should provide a publicly accessible API or downloadable software packages. Due to technological advancements, geoparsers developed before 2010 generally do not exhibit performance close to the state-of-the-art, and their source code can be hard to obtain and may not run on a modern operating systems. Third, for industrial geoparsers, they should provide an API that either allows free access or has a permissive rate limitation for free requests. These three criteria follow the foundational work of Gritta et al. (2018), and the geoparsers below are provided on EUPEG.

- **GeoTxt.** GeoTxt is an academic geoparser developed by the GeoVISTA Center at Pennsylvania State University (Karimzadeh et al., 2013; Karimzadeh, Pezanowski, MacEachren, & Wallgrün, 2019). It was initially designed to geoparse microblogs (e.g. tweets), but can be applied to longer texts as well. GeoTxt provides a publicly accessible and free API at http://www.geotxt.org, and is being maintained by its researchers. EUPEG does not host a local instance of GeoTxt but connects to its API. An advantage of such an online connection is that new updates of GeoTxt will be reflected on EUPEG. However, EUPEG cannot use GeoTxt when its online service is down. EUPEG connects to version 2.0 of GeoTxt, which uses its local GeoNames gazetteer deployed in July 2017.

- **The Edinburgh Geoparser.** The Edinburgh Geoparser is an academic geoparser developed by the Language Technology Group at the University of Edinburgh (Alex et al., 2015). A publicly available package is provided at https://www.ltg.ed.ac.uk/software/geoparser. EUPEG hosts version 1.1 of the Edinburgh Geoparser which uses the online service of GeoNames as its gazetteer. It is supported on Linux and MacOS, but cannot run on Windows.

- **TopoCluster.** TopoCluster is an academic geoparser developed by DeLozier et al. (2015) at the University of Texas at Austin. It performs geoparsing based on the geographic profiles of words characterized by the
| Data set   | Admin. units (cities, towns, ...) | Natural features (rivers, mountains, ...) | Facilities (buildings, roads, airports, ...) | Demonyms (Canadian, Syrian, American, ...) | Metonymies (London announced a new policy ...) | Modifiers (Spanish sausage, UK beef, ...) |
|------------|---------------------------------|------------------------------------------|---------------------------------------------|---------------------------------------------|-------------------------------------------------|---------------------------------------------|
| LGL        | ✓                               | ✓                                        | a                                          | ✓                                           | ✓                                               | ✓                                           |
| GeoVirus   | ✓                               |                                          |                                            |                                             |                                                 |                                             |
| TR-News    | ✓                               |                                          | a                                          |                                             |                                                 | ✓                                           |
| GeoWebNews | ✓                               | ✓                                        | ✓                                          | ✓                                           | ✓                                               | ✓                                           |
| WikToR     | ✓                               |                                          |                                            |                                             |                                                 |                                             |
| GeoCorpora | ✓                               | ✓                                        |                                            | ✓                                           | ✓                                               |                                             |
| Hu2014     | ✓                               |                                          |                                            |                                             |                                                 |                                             |
| Ju2016     | ✓                               |                                          |                                            |                                             |                                                 |                                             |

\(^a\)The data set contains toponym annotations in that category but does not provide geographic coordinates for the annotated toponyms.
local Getis-Ord $G_i^*$ statistic. While its methodology focuses on toponym resolution, its source code (https://github.com/grantdelozier/TopoCluster) provides an end-to-end system for completing both geoparsing steps. TopoCluster does not provide an official version number. We host its latest version shared on GitHub which was updated in November 2016.

- **CLAVIN.** The Cartographic Location and Vicinity Indexer is an open source geoparser that employs Apache OpenNLP Name Finder for toponym recognition, and a number of heuristics and fuzzy search for toponym resolution. CLAVIN does not come with an academic paper, but descriptions and source code can be obtained from GitHub (https://github.com/Berico-Technologies/CLAVIN) and Maven Central. We host CLAVIN 2.1.0 on EUPEG, and it employs a local GeoNames gazetteer deployed in April 2019.

- **Yahoo! PlaceSpotter.** Yahoo! PlaceSpotter is an industrial geoparser which offers an online Representational State Transfer (REST) API. As a proprietary geoparser, PlaceSpotter does not describe the exact methods employed but provides some descriptions of its functions and outputs at https://developer.yahoo.com/boss/geo/docs/PM_KeyConcepts.html. PlaceSpotter is requested via YQL (Yahoo! Query Language), and its rate limit for free requests is relatively permissive (2,000 calls per hour; a corpus with 5,000 entries can be parsed within 3 hr). EUPEG connects to Yahoo! PlaceSpotter via its online REST API which employs its Where-on-Earth ID (WOEID) for referencing places.

- **CamCoder.** CamCoder is an academic geoparser developed by Gritta et al. (2018b) at the Language Technology Lab of the University of Cambridge. CamCoder is a deep-learning-based geoparser that leverages convolutional neural networks (CNNs) with global maximum pooling and map-based word vector representations. The source code of CamCoder is available at https://github.com/milangritta/Geocoding-with-Map-Vector. Running CamCoder requires configurations on the local computing environment to include deep learning libraries, such as Tensorflow and Keras. EUPEG hosts the latest version of CamCoder shared on GitHub (updated in September 2018) which uses its local GeoNames gazetteer prepared in July 2018.

In addition to the six specialized geoparsers above, EUPEG provides three geoparsing systems extended from general NER tools (e.g. Stanford NER). These systems are included because previous research has argued that geoparsing can be considered as a subtask of NER, and a geoparser can be developed by limiting the entities recognized by an NER tool to toponyms and adding spatial footprints via a gazetteer (Inkpen, Liu, Farzindar, Kazemi, & Ghazi, 2017). Thus, including these NER-based geoparsing systems can help provide more comprehensive comparisons. The following three systems are hosted on EUPEG:

- **Stanford NER + Population.** Stanford NER is a powerful and open source NER tool developed by the Stanford Natural Language Processing Group. It has been used in numerous previous studies, including some specialized geoparsers, such as GeoTxt and TopoCluster. We extend Stanford NER to a geoparsing system by assigning the recognized toponyms to the place instances with highest populations. This simple heuristic is used because previous research has shown that assigning place names to the instances with the highest populations is a strong baseline for geoparsing and can sometimes surpass more complex models (DeLozier et al., 2015; Gritta et al., 2018a; Speriosu & Baldridge, 2013). We use Stanford CoreNLP toolbox (version 3.9.2) integrated with the online service of GeoNames.

- **spaCy NER + Population.** spaCy is a free and open source library for natural language processing tasks in Python. Released in 2014, it has already been used in many studies and applications due to its good performance (Choi, Tetreault, & Stent, 2015; Jiang, Banchs, & Li, 2016). We integrate spaCy NER (version 2.0.18) with the online service of GeoNames, and assign the recognized toponyms to the place instances with the highest populations.

- **DBpedia Spotlight.** DBpedia Spotlight is a general named entity recognition and linking (NERL) tool (Daiber, Jakob, Hokamp, & Mendes, 2013; Mendes, Jakob, García-Silva, & Bizer, 2011). This type of tool not only recognizes entities from texts but also links them to the corresponding URLs in a knowledge base (such as DBpedia) which provides geographic coordinates for the places recognized. We convert DBpedia Spotlight into a geoparser
by limiting the output to toponyms and extracting their coordinates from DBpedia pages via the geo:lat and geo:long properties. While there exist other NERL tools such as AIDA (Hoffart et al., 2011) and TagMe (Ferragina & Sciailla, 2010), DBpedia Spotlight is a widely used NERL tool whose performance compares favorably with the state-of-the-art (Cornolti et al., 2013; Usbeck et al., 2014; Van Erp, Rizzo, & Troncy, 2013). EUPEG hosts DBpedia Spotlight 1.0.0 with coordinates retrieved from online DBpedia pages.

Table 3 summarizes these systems and their main components.

### 3.4 Performance metrics

EUPEG provides a number of performance metrics based on which different geoparsers can be evaluated and compared. There is no general agreement on which metrics should be used for evaluating geoparsers. As a result, we select eight metrics that have been used in a variety of previous studies. In the following, we discuss these metrics individually.

- **Precision.** Precision (see, for example, Inkpen et al., 2017; Leidner, 2008; Lieberman et al., 2010) measures the percentage of correctly identified toponyms (true positives) among all the toponyms recognized by a geoparser, and is given by:

\[
\text{Precision} = \frac{tp}{tp + fp}
\]

where \(tp\) denotes **true positives** and \(fp\) denotes **false positives**.

- **Recall.** Recall (see, for example, Inkpen et al., 2017; Leidner, 2008; Lieberman et al., 2010) measures the proportion of correctly identified toponyms among all the toponyms that should be identified (i.e. the toponyms that are annotated as ground truth), and is given by:

\[
\text{Recall} = \frac{tp}{tp + fn}
\]

#### TABLE 3 Geoparsing systems hosted on EUPEG and their main components

| Geoparser                        | Toponym recognition | Toponym resolution | Gazetteer                          |
|----------------------------------|---------------------|--------------------|------------------------------------|
| GeoTxt (Version 2.0)             | Stanford NER        | Heuristic rules    | GeoNames (July 2017)               |
| Edinburgh (Version 1.1)          | LT-TTT2             | Heuristic rules    | GeoNames (online)                  |
| TopoCluster (Nov. 2016)          | Stanford NER        | Geoprofiles of words | GeoNames+ Natural Earth (Nov. 2016) |
| CLAVIN (Version 2.1.0)           | Apache OpenNLP      | Heuristic rules    | GeoNames (Apr. 2019)               |
| Yahoo! PlaceSpotter (online)     | Proprietary         | Proprietary        | WOEID (online)                     |
| CamCoder (Sept. 2018)            | spaCy NER           | CNNs+map-based word vectors | GeoNames (July 2018)               |
| Stanford NER + Population (Version 3.9.2) | Stanford NER | Highest population | GeoNames (online)                  |
| spaCy NER + Population (Version 2.0.18) | spaCy NER | Highest population | GeoNames (online)                  |
| DBpedia Spotlight (Version 1.0.0) | LingPipe Exact Dictionary Chunker | Context similarity | DBpedia (online)                   |
where $fn$ denotes false negatives.

- **F-score.** The F-score (see, for example, Inkpen et al., 2017; Leidner, 2008; Lieberman et al., 2010) is the harmonic mean of precision and recall. It is high when both precision and recall are fairly high and is low if either of the two is low. It is given by

$$F\text{-score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$  \hspace{1cm} (3)$$

The F-score is also called the F-measure or $F_1$-score.

- **Accuracy.** Accuracy (see, for example, Gelernter & Mushegian, 2011; Gritta et al., 2018; Karimzadeh, 2016) is suitable for measuring performances on those corpora that do not have all toponyms annotated. For example, both WikToR and Ju2016 only annotate a subset of all the toponyms mentioned in the text. In these situations, precision, recall, and F-score are no longer suitable. Accuracy can be used to quantify the proportion of the annotated toponyms that are also recognized by a geoparser. It is given by:

$$\text{Accuracy} = \frac{|\text{Annotated} \cap \text{Recognized}|}{|\text{Annotated}|}$$  \hspace{1cm} (4)$$

where Annotated denotes the set of toponyms provided in the annotation, and Recognized represents the set of toponyms recognized by the geoparser.

Precision, recall, F-score, and accuracy quantify the ability of a geoparser to correctly recognize place names from texts rather than geolocating these names. Accordingly, they measure the performance of a geoparser in the toponym recognition step. The establishment of matching between ground-truth annotations and geoparsing outputs is a topic that is worth discussing since it can directly affect the measures obtained. Previous work has discussed both exact matching and inexact matching (Gritta et al., 2018). For a sentence such as “The town of Amherst has been a leader in providing online geographic information”, a geoparser may recognize “Amherst” as a toponym, while the ground-truth annotation may be “town of Amherst”. For exact matching, this will be considered as both a false positive and a false negative, since the output of the geoparser does not match the ground truth. For inexact matching, it will be considered as a true positive. We adopt inexact matching to accommodate such syntactically inconsistent but semantically meaningful outputs, and use the same implementation as in Gritta et al. (2018) for determining matches.

To measure the performance of a geoparser in geolocating toponyms, the following four metrics are provided on EUPEG:

- **Mean error distance (MED).** The MED (see, for example, Cheng, Caverlee, & Lee, 2010; Santos et al., 2015; Speriosu & Baldridge, 2013) computes the mean of the Euclidean distances between the annotated location and the location output by a geoparser. It is given by:

$$\text{MED} = \frac{\sum_{i=1}^{N} \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}}{N}$$  \hspace{1cm} (5)$$

where $N$ is the number of annotated toponyms that are recognized and geolocated by a geoparser, $(x_i, y_i)$ are the annotated coordinates, and $(x_i', y_i')$ are the geoparsed coordinates. The toponyms, which are only in the geoparsing output or only in the annotations, are not included in computing the MED; those mismatches are evaluated by the previous four metrics.
Median error distance (MdnED). The MED is sensitive to outliers, which means a small number of geoparsed toponyms that are located far away from their ground-truth locations can greatly distort the evaluation result. The MdnED (see, for example, DeLozier et al., 2015; Santos et al., 2015; Speriosu & Baldridge, 2013) computes the median value of the error distances and is robust to outliers. It is given by:

\[
\text{MdnED} = \text{Median} \left( \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{161}\right), i \in [1,N] \]

where \(e_{di}\) denotes the \(i\)th error distance.

• Accuracy@161. This metric (see, for example, Cheng et al., 2010; DeLozier et al., 2015; Gritta et al., 2018) calculates the percentage of the toponyms that are geolocated within 161 km (100 miles) of the ground-truth locations. A main motivation for having this metric is that the geographic coordinates of a place in a gazetteer used for geoparsing may be different from the annotated coordinates. Thus, an error distance can exist even when a geoparser correctly resolves a toponym to the right place instance. Accuracy@161 considers the result as correct as long as the resolved location is within 100 miles of the annotated location. It is given by:

\[
\text{Accuracy}@161 = \frac{\left| \frac{\sqrt{(x_i - x)^2 + (y_i - y)^2}}{161}, i \in [1,N], e_{di} \leq 161 \text{ km} \right|}{N}
\]

• Area under the curve (AUC). The AUC (see, for example, Gritta et al., 2018b, 2018; Jurgens, Finethy, McCorriston, Xu, & Ruths, 2015) is a metric that quantifies the overall deviation between geoparsed locations and ground-truth annotations. It is computed by first plotting a curve of the normalized log error distance and then calculating the total area under the curve. Figure 3 shows an example of the error distance curve. The horizontal axis represents the index of the toponyms ranked from small to large error distances. A majority of toponyms are typically located at the correct locations, and therefore have zero error. However, once the error distance starts to appear, it can increase rapidly. The vertical axis represents the normalized log error distance of the geoparsed toponyms. The AUC is the total area under the curve, given by:
where $\text{Max\_Error}$ is the maximum possible error distance (half of the Earth’s circumference) between the ground truth and the geoparsed location. A better geoparser should have a lower AUC.

Calculating the error distance is a fundamental step for the four metrics above. Currently, the error distance is calculated based on point locations only. This is because all the geographic corpora we have reviewed contain only point-based annotations, and all the geoparsing systems discussed only output point-based locations. To some degree, such point-based annotation and geoparsing facilitate the comparison of geoparsing outputs and ground-truth annotations: it may be more difficult to reach an agreement on how to compare a geoparsing output that contains points, lines, and polygons with a ground-truth data set that contains, say, points only. However, geolocating a toponym to a single point is not ideal when the toponym refers to a large geographic area (e.g. a country or a river). Geographic scale further complicates this issue: it may be fine to locate a city name at a point if a study focuses on the country scale, but we may want a city name or even a neighborhood name to be represented as a polygon if the study is at the city scale. While existing geoparsers only output point-based locations, future geoparsers could provide other geometries to represent the spatial footprints of the toponyms recognized. When geoparsers and annotated data sets with various footprints have become available, EUPEG could be extended with error distances calculated using other methods such as Fréchet distance and Hausdorff distance.

In sum, EUPEG provides eight metrics for quantifying the performance of geoparsers. Four of these metrics examine the percentage of the toponyms correctly recognized from texts, while the other four metrics are based on the error distances between the geoparsed locations and their ground-truth locations. There also exist ranking-based metrics, such as mean reciprocal rank and normalized discounted cumulative gain (Purves et al., 2018). However, these metrics require a geoparser to output a ranked list of candidate places. Some end-to-end systems only output one single result rather than a list of places. Thus, these ranking-based measures are not included.

3.5 Resource unification

The corpora and geoparsers hosted on EUPEG are from different sources and are highly heterogeneous. They can be annotated using different data formats. For example, LGL, GeoVirus, TR-News, GeoWebNews, and WikToR use Extensible Markup Language (XML), and organize data into hierarchical structures; Hu2014 and Ju2016 use TXT or comma-separated values (CSV), and organize data using simple line-by-line text annotations with each line representing one data record; GeoCorpora uses the format of tab-separated values (TSV), and one data record can be put on multiple lines if it contains more than one toponym. The outputs of geoparsers are also in different formats. GeoTxt and Yahoo! PlaceSpotter use JavaScript Object Notation (JSON) to format their outputs; the Edinburgh Geoparser employs XML; TopoCluster uses its own text-based geoparsing output; and CLAVIN provides an API that allows a user to format the output in a customized manner. Even if the resources are in the same format, they can still use different vocabularies to organize similar content (which will be discussed in Section 3.6).

EUPEG serves as a platform for unifying these heterogeneous geoparsing resources. Building on the foundational work of Gritta et al. (2018), we unify these resources in the following steps. First, we write a customized computer script for each geographic corpus to convert it into two parts: a collection of individual text files (with each file containing one text entry) and a single ground-truth text file (with each line containing the ground-truth annotation for one file in the collection). Such a design was used in Gritta et al. (2018). While it seems to be rather an engineering design, we reuse it since EUPEG is built on the work of Gritta et al. (2018) and doing so avoids reinventing the wheel. Second, we write a customized wrapper for each geoparser hosted on EUPEG. These wrappers convert the heterogeneous geoparsing outputs into the same format in which each line contains the recognized toponyms from one text file. Third, a comparison function is developed to compare the standardized geoparsing...
outputs with the ground-truth files, and measure the performances of the geoparsers by computing the eight metrics. In sum, EUPEG unifies the heterogeneous resources by first converting them into the same formats and then comparing the performances of geoparsers based on the same metrics.

3.6 Resource extension

The resources on EUPEG can be extended with new corpora and geoparsers. A newly created geographic corpus can be uploaded to EUPEG to test the performance of geoparsers. To enable the upload of any new corpus to EUPEG, we need an agreed format for organizing the text entries and ground-truth annotations in the new corpus. Although some toponym annotation languages, such as TRML (Leidner, 2006) and SpatialML (Mani et al., 2010), have been proposed, many publicly shared corpora, such as LGL and WikToR, use their own formats, probably due to a lack of access to example data sets of TRML and SpatialML. Here, we specify the format of a new corpus to be connected to EUPEG based on LGL, GeoVirus, TR-News, GeoWebNews, and WikToR. Although these five corpora are all in XML format, they employ different XML tags to organize their content. For example, GeoVirus uses the XML tag `<article>` to represent each text entry, while WikToR uses the tag `<page>` to represent each entry (since the text entries are Wikipedia pages). Similarly, TR-New uses the tag `<gaztag>` to provide location information obtained from a gazetteer, while GeoWebNews does not use the tag `<gaztag>` at all. Learning from these existing corpora, we build an XML format that has a small number of required core tags and offers the flexibility of including optional tags. Listing 1 shows this format.

```xml
<?xml version="1.0" encoding="utf-8"?>
<entries>
  <entry>
    <text>Paris is a city in Texas...</text>
    <toponyms>
      <toponym>
        <start>0</start>
        <end>4</end>
        <phrase>Paris</phrase>
        <place>
          <footprint>-95.5477 33.6625</footprint>
          <placename>City of Paris</placename>  # optional
          <placetype>ADM3</placetype>  # optional
        </place>
      </toponym>
      ...  # other optional attributes
    </toponyms>
    <toponym>
      ...  # another annotated toponym
    </toponym>
    ...
  </entry>
  <entry>
    ...  # another entry in the dataset
  </entry>
  ...
</entries>
```

**LISTING 1** The format for a new corpus to be uploaded to EUPEG
A corpus to be uploaded to EUPEG will be organized into a single XML file using the format above. This file can contain multiple text \(<\text{entries}\)> entries, and the \(<\text{entry}\)> tag is used to organize each individual data entry. The \(<\text{text}\)> tag contains the text to be geoparsed, which can be a news article, a tweet, or a web page, for example. The \(<\text{toponyms}\)> tag contains the toponyms in the ground-truth annotation. Each ground-truth \(<\text{toponym}\)> should contain the \(<\text{start}\)> position (in character index) and the \(<\text{end}\)> position of the toponym in the text. The \(<\text{phrase}\)> tag contains the name of the place mentioned in the text, which can be not only an official name but also a name abbreviation, a colloquial name, or other alias. The \(<\text{place}\)> tag contains the annotated information about the place. The required core information for a \(<\text{place}\)> is \(<\text{footprint}\>) in the form of longitude and latitude. Other optional information, such as \(<\text{placename}\>) and \(<\text{placetype}\>) can also be included.

A newly developed geoparser can be connected to EUPEG and compared with other hosted geoparsers. To do so, one needs to make the new geoparser accessible via a REST API and organize the geoparsing output using an agreed format. As far as we know, there is no standard way to organize geoparsing output. Accordingly, we specify the output format based on that of an existing geoparser, GeoTxt. As an academic geoparser, GeoTxt is available freely and publicly with a REST API and accompanied by scholarly publications (Karimzadeh et al., 2013, 2019). GeoTxt uses JSON to format its output. The original output of GeoTxt contains elements specific to the gazetteer used, GeoNames, such as \(<\text{geoNameId}\>) and \(<\text{featureCode}\>) . These elements are not required in this format since a new geoparser may not necessarily employ GeoNames as its gazetteer. Similarly to the format of a new corpus, we also classify the information elements in the geoparsing output as required core elements and optional ones. A geoparser can output only the four core elements for simple implementation, or can include additional and optional information for a comprehensive output. The proposed output format for a new geoparser is shown in Listing 2.

This format organizes the geoparsing output into a JSON object. It starts with a root attribute \(<\text{toponyms}\>) whose value is an array of JSON objects. Each JSON object contains the information for a toponym recognized by the geoparser. The attribute \(<\text{start}\>) contains the start position (in character index) of the toponym, while

```json
{
  toponyms: [
    {
      start: 0,
      end: 4,
      phrase: "Paris",
      place: {
        footprint: [[-95.5477, 33.6625]],
        placename: "City of Paris", # optional
        placetype: "ADM3" # optional
      },
      ... # other optional attributes
    },
    {
      ... # another recognized toponym
    },
    ...
  ]
}
```

**Listing 2** The format for the output of a new geoparser to be connected to EUPEG
the attribute end contains the end position of the recognized toponym. The attribute phrase represents the toponym mentioned in the text, which could be an official name or other alternative name. The attribute place contains more detailed information about the place recognized. The required element is footprint, which takes the value of a JSON array following the format of GeoJSON. For a typical point-based footprint, the JSON array contains the longitude and latitude of the place. Other optional information, such as placename and placetype, can also be included.

3.7 | Experiment archiving and search

Another important function of EUPEG is archiving experiments. A database is created to store information about an experiment, such as experiment ID, date and time, data sets, geoparsers, metrics, and experiment results. An experiment ID is a 16-digit serial number that uniquely identifies an experiment. All other information is based on the configurations specified by a user at the time of running an experiment. One can search experiments based on their IDs and see their results.

The value of this function can be seen in two aspects. First, it facilitates the sharing of experiment results. A researcher or a geoparser user can quickly share the result of an experiment with colleagues by embedding the experiment ID in an email, for example. The colleagues who receive this experiment ID can check it on EUPEG and see the experiment results and configurations themselves. Second, the independently recorded experiment results provide further evidence for researchers to demonstrate their work, and allow others to verify the outcome of a study more easily. Accordingly, EUPEG can help enhance the reproducibility and replicability of scientific research.

3.8 | Summary

We have presented the overall architecture, resources, and functions of EUPEG. In summary, EUPEG has the following features:

- **Comprehensiveness.** EUPEG provides eight annotated corpora, nine geoparsing systems, and eight performance metrics for evaluating geoparsers. The annotated corpora are in four different text genres; the geoparsing systems include both specialized geoparsers and those extended from general named entity recognizers; and the performance metrics include metrics based on information retrieval and those based on error distances.
- **Unification.** EUPEG can be considered as a one-stop platform where corpora, geoparsers, and performance metrics are unified. EUPEG also unifies geoparser users and geoparser researchers: users can use EUPEG to select the most suitable geoparser for their own corpora, while researchers can leverage the resources hosted to perform effective and efficient evaluation experiments.
- **Extensibility.** EUPEG offers extensibility for the geoparsing resources hosted. A newly created corpus can be uploaded to EUPEG to test the geoparsers hosted. A newly developed geoparser can be connected to EUPEG and compared with other geoparsers. We also provide the EUPEG source code, and EUPEG can be further extended by adding new performance metrics or other features for evaluating geoparsers.
- **Documentation.** EUPEG documents experiment results and configurations, and provides a search function for retrieving previous experiments. Such an archiving feature provides researchers with further evidence to demonstrate their research outcome. It also enables researchers and users to share experiment results more easily by, for example, embedding the experiment ID in an email.
4 | IMPLEMENTATION AND ANALYTICAL EVALUATION

4.1 | Implementation and demonstration

Based on the proposed architecture, we have implemented EUPEG as a web-based platform that can be accessed online at https://geoai.geog.buffalo.edu/EUPEG. Figure 4 shows a screenshot of its main interface. EUPEG offers a (1)–(2)–(3) workflow for conducting an experiment: a user selects (1) data sets, (2) geoparsers, and (3) metrics, and then clicks the “Run this experiment” button to start the experiment (Figure 4). Users can also click the “Add corpus...” or “Add geoparser...” buttons to add their own resources. Once an experiment is finished, the user will be provided with an experiment ID which can then be used by the user or others to search for the experiment. Figure 5 shows an example of searching a previous experiment and seeing its results. The results returned contain not only the performance information of the geoparsers compared based on the selected corpora and metrics, but also the date and time of the experiment, the versions of the geoparsers, and the gazetteers used. Such information allows the user to see the detailed configuration of an experiment. The reader can also try out the platform by searching for the experiment ID “8380NII17XEKM0GD” on EUPEG.

EUPEG is implemented using a technology stack of multiple programming languages, software libraries, and development tools. Java JDK 11 is used on the server side to implement servlets, database connections, and external API requests. Javascript, HTML5, CSS3, and other libraries, such as Bootstrap and JQuery, are employed on the client side to implement the user interface and AJAX-based HTTP requests and responses. SQLite 3 is used to store the experiment records; this is a light-weight, high-reliability, and public-domain database. To reduce the time of experiments and avoid running the same experiment many times, we store and reuse the results if the data sets and geoparsers selected by a user were tested in a previous experiment. Such an implementation increases experiment efficiency and decreases computation costs, since running an experiment can take from hours to days depending on the data sets and geoparsers selected. In addition, we use the following approaches to keep the hosted geoparsers up-to-date. For the geoparsers that are connected to EUPEG via online APIs, a computational thread is developed which runs in the background and automatically updates the geoparsing results of these...
geoparsers once per month to reflect any possible changes. For the geoparsers that are deployed locally on our server, we plan to check their websites once every 3–6 months and will update our local instances when new stable versions become available. While we plan to maintain EUPEG for the next few years, resource limitations may not allow us to maintain it for a long time. Thus, we also share the source code of EUPEG, along with the data sets under permitted licenses (e.g. GNU General Public License), on GitHub at https://github.com/geoai-lab/EUPEG, and invite the community to further enhance and extend it.

4.2 Analytical evaluation

An important aim of EUPEG is to reduce the time that researchers have to spend in preparing data sets and baselines for experiments. This section attempts to estimate the amount of time that could be saved by EUPEG. One possible approach to providing such an estimate would be to ask a number of researchers to prepare all the corpora and geoparsers hosted on EUPEG by themselves, and measure the average time they spend. Such a process, however, could be very tedious for the researchers concerned and, depending on their particular fields and technical skills, the time they take might not be representative of the time that others might need to prepare these experiment resources. Here, we provide an analytical evaluation of the amount of time that could be saved, based on our own experience of developing EUPEG, and focus on the lower bound of the time. In the following, we first analyze the steps that a research group typically has to complete in order to prepare data sets and baselines for an experiment themselves. These steps are shown in Figure 6. We then estimate the minimum amount of time that is necessary to complete each step.

**FIGURE 5** Display of the results of a previous experiment
Read literature and identify resources. This is generally the first step, in which one studies previous research and identifies resources that can be reused. For this step, EUPEG does not save much time. Although it makes various resources ready for use, researchers may still need to read the related publications to understand the methods under the hood. EUPEG and this article, however, can serve as an entry point for new researchers. The time that can be saved in this step is estimated as zero.

Search, find, and download resources. After identifying resources from the literature, one needs to obtain them. For most data sets hosted on EUPEG, we were able to obtain each of them within half an hour, thanks to the authors who shared relevant URLs in their papers. For GeoCorpora, while the authors have kindly provided its URL, much time is still needed to rehydrate this data set due to Twitter’s data sharing restrictions. It took us more than 5 person-hours to recover this data set, and additional time has to be spent on applying for a Twitter developer account before one can start to recover the data set. We estimate a minimum of 8.5 person-hours for obtaining the data sets hosted on EUPEG. For the geoparsers, we were able to download the source codes or compiled versions of the Edinburgh Geoparser, TopoCluster, CLAVIN, and CamCoder within half an hour each. The other five geoparsers are either connected to EUPEG via their APIs or are further developed based on general NER tools. About half an hour is needed to find each of these resources. In total, about 13 person-hours are needed for this step.

Read and understand documentation. After the source codes of previous geoparsers are obtained, one needs to read documentation and understand how to deploy and run them. Background knowledge of different programming languages (e.g. Python and Java) and system architectures (e.g. REST web services) is necessary to understand the installation instructions. Based on our own experience, we estimate an average of 2 person-hours for an experienced developer to read and understand the documentation of one geoparsing system. Thus, this step takes about 18 person-hours.

Deploy and configure geoparsers. This step is particularly time-consuming and requires a lot of expertise. First, different geoparsers can be implemented in different programming languages. Accordingly, a researcher needs to have some basic knowledge of the various languages in order to deploy them. Second, there exist specific configuration requirements for some geoparsers. For example, geoparsers available via web APIs require one to have expertise in handling HTTP requests and responses; a geoparser (e.g. TopoCluster) may require the installation of a database and its spatial extension, or may require a researcher to be familiar with certain deep learning libraries (e.g. CamCoder). Third, including general NER tools as baselines requires further developments and gazetteer configurations to convert these general tools into geoparsers. We estimate an average of 24 person-hours to successfully deploy and configure one geoparser, and thus a total of 216 person-hours.

Harmonize the formats of resources. The annotated data sets and the outputs of geoparsers are often in different formats and structures. To conduct an experiment on these heterogeneous resources, one needs to harmonize these data sets and geoparser outputs by writing programs to convert them into the same format. We estimate an
average of 3 person-hours to process one resource (a data set or a geoparser), and in total, this step takes about 51 person-hours.

*Implement performance metrics.* Performance metrics, such as precision, recall, and the AUC, need to be implemented in order to evaluate geoparsers. In addition, some programming work is necessary to be able to compare geoparsing outputs to ground-truth annotations. In total, we estimate 11 person-hours to complete this step (8 hr for eight metrics plus 3 hr to develop the code for comparing outputs with ground-truth annotations).

*Run experiments with resources.* Once everything is prepared, we can run experiments to obtain evaluation results. The running time of different geoparsers can vary largely. Figure 7 reports the empirical time of the nine geoparsers on the same machine for processing the GeoCorpora data set. TopoCluster took the longest time (660.51 min), while CLAVIN is the fastest geoparser, taking only 0.28 min to process the same data set. A longer processing time, however, does not mean better performance. Figure 8 shows the performance of different geoparsers on GeoCorpora. The time that can be saved by EUPEG in this step is estimated as zero, since one can work on other tasks when an experiment is running.

**FIGURE 7** Running time of different geoparsers on GeoCorpora

| Geoparser Name   | Precision | Recall | F-Score | Accuracy | Mean (km) | Median (km) | Accuracy@161 | AUC   |
|------------------|-----------|--------|---------|----------|-----------|-------------|---------------|-------|
| GeoTxt           | 0.978     | 0.550  | 0.704   | 0.550    | 786.140   | 0.000       | 0.848         | 0.137 |
| Edinburgh        | 0.888     | 0.539  | 0.671   | 0.539    | 1179.604  | 0.000       | 0.803         | 0.176 |
| TopoCluster      | 0.950     | 0.545  | 0.693   | 0.545    | 746.734   | 38.764      | 0.670         | 0.381 |
| CLAVIN           | 0.966     | 0.363  | 0.528   | 0.363    | 609.056   | 0.000       | 0.858         | 0.131 |
| Yahoo! PlaceSpotter | 0.871   | 0.640  | 0.738   | 0.640    | 564.471   | 49.980      | 0.681         | 0.390 |
| CamCoder         | 0.931     | 0.518  | 0.666   | 0.518    | 1095.508  | 0.000       | 0.790         | 0.186 |
| StanfordNER      | 0.970     | 0.568  | 0.716   | 0.568    | 1269.950  | 0.456       | 0.649         | 0.296 |
| SpaCyNER         | 0.799     | 0.530  | 0.637   | 0.530    | 1372.139  | 0.000       | 0.690         | 0.279 |
| DBpedia Spotlight| 0.912     | 0.526  | 0.667   | 0.526    | 760.484   | 33.816      | 0.646         | 0.356 |

**FIGURE 8** The performance of different geoparsers on GeoCorpora
Summarize experiment results. When experiments are finished, one often needs to collect the results and organize them into a report. For this step, we estimate the saved time as zero, since it can be done with or without EUPEG.

Table 4 provides a summary of the approximate number of person-hours that can be saved by EUPEG.

In total, we estimate 309 person-hours if another research group were to prepare the same resources hosted on EUPEG. This estimate is close to a lower bound, as it is based on the assumption that researchers have all the necessary knowledge and technical skills and does not take into account the time spent on trial and error.

| Task for preparing experiments                              | Estimated time (person-hours) |
|--------------------------------------------------------------|-------------------------------|
| Read literature and identify resources                       | 0                             |
| Search, find, and download resources                        | 13                            |
| Read and understand documentation                           | 18                            |
| Deploy and configure geoparsers                             | 216                           |
| Harmonize the formats of resources                          | 51                            |
| Implement performance metrics                               | 11                            |
| Run experiments with resources                               | 0                             |
| Summarize experiment results                                | 0                             |
| Total                                                        | 309                           |

5 | CONCLUSIONS AND FUTURE WORK

In this work we present EUPEG, an Extensible and Unified Platform for Evaluating Geoparsers. With large amounts of textual data available from various sources, geoparsers have become increasingly important, given their capabilities of extracting geographic information from textual documents. Many studies in spatial data science and digital humanities have leveraged geoparsers in various contexts (e.g. disaster responses, spatial studies, and event detection) to integrate spatial and textual analysis. While a number of geoparsers have been developed, they have been tested on different data sets using different performance metrics. Consequently, the evaluation results reported cannot be directly compared. In addition, a new geoparser often needs to be compared with existing baselines to demonstrate its merits. However, preparing baselines and testing data sets can take much time and effort. In this context, we propose and develop EUPEG as a benchmarking platform for evaluating geoparsers and eventually enhancing spatial and textual analysis. It is implemented as a web-based and open source platform with four major features. First, it is comprehensive: EUPEG provides eight open corpora, nine geoparsing systems, and eight performance metrics for evaluating geoparsers. Second, it is unified: EUPEG can be considered as a one-stop platform consisting of heterogeneous corpora, geoparsers, and metrics. Third, it is extensible: EUPEG allows the hosted resources to be extended with new corpora and geoparsers. Its final feature is its documentation: EUPEG documents experiment results and configurations, and allows the search of previous experiments.

The main aim of EUPEG is to enable effective and efficient comparisons of geoparsers while reducing the time that researchers and end users have to spend preparing data sets and baselines. Based on our analytical evaluation, EUPEG can save one single research group approximately 309 hr in preparing the same data sets, geoparsers, and metrics. The number of hours saved will be multiplied when multiple research groups attempt to develop and compare geoparsers. While EUPEG serves as a benchmarking platform, it is not intended to replace...
project-specific evaluations necessary for highlighting certain unique features of a geoparser but to supplement existing evaluations.

The development of EUPEG also reveals several issues that may need future work. First, there is a lack of commonly agreed standards on corpus annotation. While languages, such as TRML (Leidner, 2006) and SpatialML (Mani et al., 2010), have been proposed, they have not been adopted by the recently available and publicly shared corpora, such as LGL, WikToR, and GeoCorpora. Having a commonly agreed standard can facilitate the development of data sets that are more readily usable by others in future experiments. Second, a similar situation holds for the outputs of geoparsers, where a commonly agreed output format is not available. Most geoparsers organize their outputs in a format that they consider suitable. While it is feasible to harmonize these heterogeneous outputs by writing wrapper programs (as done in EUPEG), a standard output format could make it easier for others to use a geoparser or to combine multiple geoparsers. In this work we have developed a simple format based on GeoTxt to allow new geoparsers to be connected to EUPEG. However, further efforts are needed from the community to develop an agreed and standard output format for geoparsers. Third, the current version of EUPEG focuses on English-based geoparsers and corpora only. Resources for other languages could be added in the future to support multilingual geoparsing evaluations. With the source code shared, new extensions could be added to EUPEG to further enhance it and help it better serve our community.

ACKNOWLEDGMENTS
The authors would like to thank Drs. Morteza Karimzadeh and Alan MacEachren for providing further technical details about GeoTxt.

ORCID
Yingjie Hu https://orcid.org/0000-0002-5515-4125

REFERENCES
Adams, B., & McKenzie, G. (2013). Inferring thematic places from spatially referenced natural language descriptions. In D. Sui, S. Elwood, & M. F. Goodchild (Eds.), Crowdsourcing geographic knowledge: Volunteered geographic information (VGI) in theory and practice (pp. 201–221). Berlin, Germany: Springer.
Alex, B., Byrne, K., Grover, C., & Tobin, R. (2015). Adapting the Edinburgh geoparser for historical georeferencing. International Journal of Humanities & Arts Computing, 9(1), 15–35.
Amitay, E., Har’El, N., Sivan, R., & Soffer, A. (2004). Web-a-where: Geotagging web content. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 273–280). Sheffield, UK: ACM.
Ballatore, A., & Adams, B. (2015). Extracting place emotions from travel blogs. In Proceedings of the 18th AGILE Conference on Geographic Information Science. Lisbon, Portugal: AGILE.
Barbaresi, A. (2017). Towards a toolbox to map historical text collections. In Proceedings of the 11th Workshop on Geographic Information Retrieval. Heidelberg, Germany: ACM.
Cano, A. E., Rizzo, G., Varga, A., Rowe, M., Stankovic, M., & Dadzie, A.-S. (2014). Making sense of microposts: (#Microposts2014) named entity extraction & linking challenge. In M. Rowe, M. Stakovic, & A.-S. Dadzie (Eds.), Proceedings of the 4th Workshop on Making Sense of Microposts (CEUR Workshop Proceedings, Vol. 1141, pp. 54–60). Retrieved from http://ceur-ws.org/Vol-1141/microposts2014_proceedings.pdf
Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (pp. 759–768). Toronto, Canada: ACM.
Choi, J. D., Tetreault, J., & Stent, A. (2015). It depends: Dependency parser comparison using a web-based evaluation tool. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Vol. 1, Long Papers, pp. 387–396). Beijing, China: ACL.
Cornolti, M., Ferragina, P., & Ciaramita, M. (2013). A framework for benchmarking entity annotation systems. In Proceedings of the 22nd International Conference on World Wide Web (pp. 249–260). Rio de Janeiro, Brazil: ACM.

Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In Proceedings of the 9th Conference on Semantic Systems (pp. 121–124). Graz, Austria: ACM.

DeLozier, G., Baldridge, J., & London, L. (2015). Gazetteer-independent toponym resolution using geographic word profiles. In Proceedings of the 29th Conference on Artificial Intelligence (pp. 2382–2388). Palo Alto, CA: AAAI.

DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016). Creating a novel geolocation corpus from historical texts. In Proceedings of the 10th Linguistic Annotation Workshop (pp. 188–198). Berlin, Germany: ACL.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). Miami, FL: IEEE.

Faulconbridge, J. R., Hall, S. J., & Beaverstock, J. V. (2008). New insights into the internationalization of producer services: Organizational strategies and spatial economies for global headhunting firms. Environment & Planning A, 40(1), 210–234.

Ferragina, P., & Scaiella, U. (2010). Tagme: On-the-fly annotation of short text fragments (by Wikipedia entities). In Proceedings of the 19th ACM International Conference on Information and Knowledge Management (pp. 1625–1628). Toronto, Canada: ACM.

Frank, J. R., Rauch, E. M., & Donoghue, K. (2006). Spatially coding and displaying information. US Patent 7,117,199. Alexandria, VA: US Patent and Trademark Office.

Freire, N., Borbinha, J., Calado, P., & Martins, B. (2011). A metadata geoparsing system for place name recognition and resolution in metadata records. In Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (pp. 339–348). Ottawa, Canada: ACM.

Gao, S., Janowicz, K., Montello, D. R., Hu, Y., Yang, J.-A., McKenzie, G., ... Yan, B. (2017). A data-synthesis-driven method for detecting and extracting vague cognitive regions. International Journal of Geographical Information Science, 31(6), 1245–1271.

Geiß, J., Spitz, A., Strötgen, J., & Gertz, M. (2015). The Wikipedia location network: Overcoming borders and oceans. In Proceedings of the 7th Workshop on Geographic Information Retrieval. Paris, France: ACM.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. GeoInformatica, 17(4), 635–667.

Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. Transactions in GIS, 15(6), 753–773.

Gritt, M., Pilehvar, M. T., & Collier, N. (2018a). A pragmatic guide to geoparsing evaluation. Preprint, arXiv:1810.12368.

Gritt, M., Pilehvar, M. T., & Collier, N. (2018b). Which Melbourne? Augmenting geocoding with maps. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1, Long Papers, pp. 1285–1296). Melbourne, Australia: ACL.

Gritt, M., Pilehvar, M. T., Limspatham, N., & Collier, N. (2018). What’s missing in geographical parsing? Language Resources & Evaluation, 52(2), 603–623.

Grossner, K., Janowicz, K., & Keßler, C. (2016). Place, period, and setting for linked data gazetteers. In M. L. Berman, R. Mostern, & H. Southall (Eds.), Placing names: Enriching and integrating gazetteers (pp. 80–96). Indianapolis: Indiana University Press.

Hecht, B., & Moxley, E. (2009). Terabytes of Tobler: Evaluating the first law in a massive, domain-neutral representation of world knowledge. In K. S. Hornsby, C. Claramunt, M. Denis, & G. Ligozat (Eds.), Spatial information theory: COSIT 2009 (Lecture Notes in Computer Science, Vol. 5756, pp. 88–105). Berlin, Germany: Springer.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., ... Weikum, G. (2011). Robust disambiguation of named entities in text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 782–792). Edinburgh, Scotland: ACL.

Hu, Y. (2018). EUPEG: Towards an extensible and unified platform for evaluating geoparsers. In Proceedings of the 12th Workshop on Geographic Information Retrieval (pp. 3.1–3.2). Seattle, WA: ACM.

Hu, Y., Janowicz, K., & Prasad, S. (2014). Improving Wikipedia-based place name disambiguation in short texts using structured data from DBpedia. In Proceedings of the Eighth Workshop on Geographic Information Retrieval. New York, NY: ACM.

Hu, Y., Ye, X., & Shaw, S.-L. (2017). Extracting and analyzing semantic relatedness between cities using news articles. International Journal of Geographical Information Science, 31(12), 2427–2451.

Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., & Ghazi, D. (2017). Location detection and disambiguation from twitter messages. Journal of Intelligent Information Systems, 49(2), 237–253.

Jiang, R., Banchs, R. E., & Li, H. (2016). Evaluating and combining name entity recognition systems. In Proceedings of the Sixth Named Entity Workshop (pp. 21–27). Berlin, Germany: ACL.

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. International Journal of Geographical Information Science, 22(3), 219–228.
Jones, C. B., Purves, R. S., Clough, P. D., & Joho, H. (2008). Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science*, 22(10), 1045–1065.

Ju, Y., Adams, B., Janowicz, K., Hu, Y., Yan, B., & McKenzie, G. (2016). Things and strings: Improving place name disambiguation from short texts by combining entity co-occurrence with topic modeling. In E. Blomqvist, P. Ciancarini, F. Poggi, & F. Vitali (Eds.), *Knowledge engineering and knowledge management: EKAω 2016* (Lecture Notes in Computer Science, Vol. 10024, pp. 353–367). Cham, Switzerland: Springer.

Jurgens, D., Finethy, T., McCroriston, J., Xu, Y. T., & Ruths, D. (2015). Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media* (pp. 188–197). Oxford, UK: AAAI.

Kamaloo, E., & Rafiei, D. (2018). A coherent unsupervised model for toponym resolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web* (pp. 1287–1296). Lyon, France: ACM.

Karimzadeh, M. (2016). Performance evaluation measures for toponym resolution. In *Proceedings of the 10th Workshop on Geographic Information Retrieval*. Burlingame, CA: ACM.

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., … MacEachren, A. M. (2013). Geotxt: A web API to leverage place references in text. In *Proceedings of the Seventh Workshop on Geographic Information Retrieval* (pp. 72–73). Orlando, FL: ACM.

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.

Ladra, S., Luaces, M. R., Pedreira, O., & Seco, D. (2008). A toponym resolution service following the OGC WPS standard. In M. Bertolotto, C. Ray, & X. Li (Eds.), *Web and wireless geographical information systems: W2GIS 2008* (Lecture Notes in Computer Science, Vol. 5373, pp. 75–85). Berlin, Germany: Springer.

Lan, R., Adelfio, M. D., & Samet, H. (2014). Spatio-temporal disease tracking using news articles. In *Proceedings of the Third ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health* (pp. 31–38). Dallas, TX: ACM.

Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment & Urban Systems*, 30(4), 400–417.

Leidner, J. L. (2008). *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Irvine, CA: Universal Publishers.

Li, H., Srihari, R. K., Niu, C., & Li, W. (2002). Location normalization for information extraction. In *Proceedings of the 19th International Conference on Computational Linguistics* (Vol. 1). Taipei, Taiwan: ACL.

Lieberman, M. D., Samet, H., & Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *Proceedings of the 26th IEEE International Conference on Data Engineering* (pp. 201–212). Long Beach, CA: IEEE.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing relatedness by toponym co-occurrences on web pages. *Transactions in GIS*, 18(1), 89–107.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., … Blanford, J. (2011). Senseplace2: Geotwitter analytics support for situational awareness. In *Proceedings of the 2011 IEEE Conference Visual Analytics Science and Technology* (pp. 181–190). Providence, RI: IEEE.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., … Clancy, S. (2010). SpatialML: Annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3), 263–280.

Mani, I., Hitzeman, J., Richer, J., & Harris, D. (2008). *ACE 2005 English SpatialML annotations*. Philadelphia, PA: Linguistic Data Consortium.

McKenzie, G., Liu, Z., Hu, Y., & Lee, M. (2018). Identifying urban neighborhood names through user-contributed online property listings. *ISPRS International Journal of Geo-Information*, 7(10), 388.

Melo, F., & Martins, B. (2017). Automated geocoding of textual documents: A survey of current approaches. *Transactions in GIS*, 21(1), 3–38.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the Seventh International Conference on Semantic Systems*. Graz, Austria: ACM.

Monteiro, B. R., Davis, C. A., Jr., & Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96, 23–34.

Nesi, P., Pantaleo, G., & Tenti, M. (2016). Geographical localization of web domains and organization addresses recognition by employing natural language processing, pattern matching and clustering. *Engineering Applications of Artificial Intelligence*, 51, 202–211.

Overell, S., & Rüger, S. (2008). Using co-occurrence models for placename disambiguation. *International Journal of Geographical Information Science*, 22(3), 265–287.

Pezanowski, S., MacEachren, A. M., Savelyev, A., & Robinson, A. C. (2018). SensePlace3: A geovisual framework to analyze place-time-attribute information in social media. *Cartography & Geographic Information Science*, 45(5), 420–437.
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.

Richter, L., Geiß, J., Spitz, A., & Gertz, M. (2017). HeidelPlace: An extensible framework for geoparsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 85–90). Copenhagen. Denmark: ACL.

Salvini, M. M., & Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multi-relational world city network. *Environment & Planning B*, 43(1), 228–248.

Santos, J., Anastácio, I., & Martins, B. (2015). Using machine learning methods for disambiguating place references in textual documents. *GeoJournal*, 80(3), 375–392.

Speriosu, M., & Baldridge, J. (2013). Text-driven toponym resolution using indirect supervision. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (pp. 1466–1476). Sofia, Bulgaria: ACL.

Sundheim, B. M. (1993). Tipster/MUC-5: Information extraction system evaluation. In *Proceedings of the Fifth Conference on Message Understanding* (pp. 27–44). Baltimore, MD: ACL.

Tjong Kim Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning* (Vol. 4, pp. 142–147). Edmonton, Canada: ACL.

Usbeck, R., Ngonga Ngomo, A.-C., Röder, M., Gerber, D., Coelho, S. A., Auer, S., & Both, A. (2014). Agdistis-graph-based disambiguation of named entities using linked data. In P. Mika, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, ... C. Goble (Eds.), *The Semantic Web - ISWC 2014: 13th International Semantic Web Conference, Riva Del Garda, Italy, October 19–23, 2014, Proceedings* (pp. 457–471). Berlin, Germany: Springer.

Usbeck, R., Röder, M., Ngonga Ngomo, A.-C., Baron, C., Both, A., Brümmer, M., ... Wesemann, L. (2015). GERBIL: General entity annotator benchmarking framework. In *Proceedings of the 24th International Conference on World Wide Web* (pp. 1133–1143). Florence, Italy: ACM.

Van Erp, M., Rizzo, G., & Troncy, R. (2013). Learning with the web: Spotting named entities on the intersection of nerd and machine learning. In *Proceedings of the Third Workshop on Making Sense of Microposts* (pp. 27–30). Rio de Janeiro, Brazil: ACM.

Wallgrün, J. O., Karimzadeh, M., MacEachren, A. M., & Pezanowski, S. (2018). GeoCorpora: Building a corpus to test and train microblog geoparsers. *International Journal of Geographical Information Science*, 32(1), 1–29.

Woodruff, A. G., & Plaunt, C. (1994). GIPSY: Automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9), 645–655.

---

**How to cite this article:** Wang J, Hu Y. Enhancing spatial and textual analysis with EUPEG: An extensible and unified platform for evaluating geoparsers. *Transactions in GIS*. 2019;23:1393–1419. [https://doi.org/10.1111/tgis.12579]