HeidelPlace: An Extensible Framework for Geoparsing

Ludwig Richter, Johanna Geiß, Andreas Spitz and Michael Gertz

Database Systems Research Group, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

What is Geoparsing? Why HeidelPlace?

Motivation: Geoparsing is a key task in text processing and central to subsequent spatial analyses. It describes the process of identifying place mentions (topoynms) and linking them to unambiguous spatial references. For example, geoparsing *"Heidelberg was founded in 1196 AD"* may reveal a reference to the German town Heidelberg.

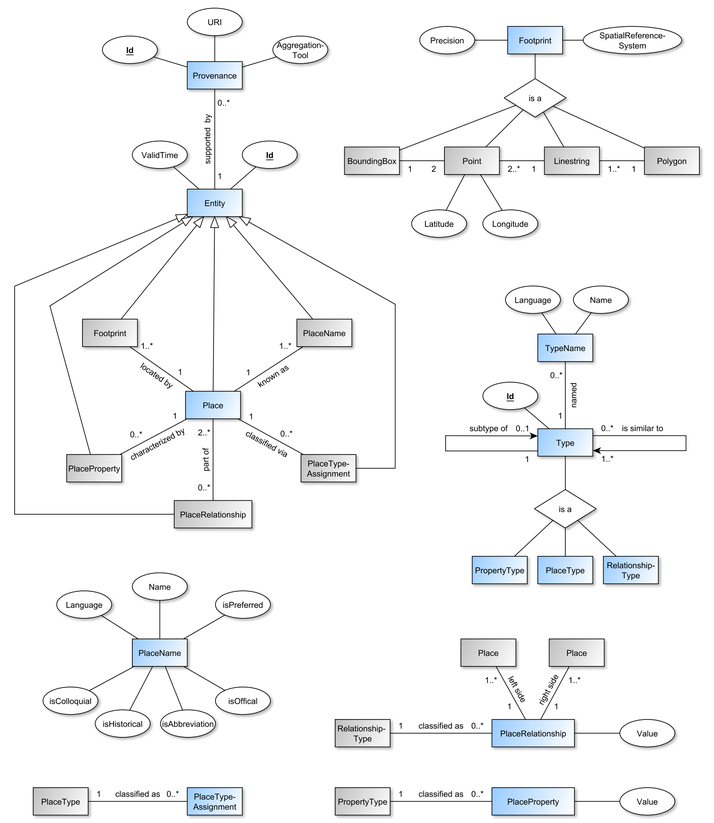
Problem: Several geoparsers are available, each with their own gazetteer and toponym recognition and resolution approaches. However, they often lack extensibility, implementations are not accessible, or they are fixed to a particular gazetteer. This makes adjustments to other application domains difficult and prevents easy experimental setup.

HeidelPlace provides:

- A generic gazetteer model supporting integration of place information from heterogeneous knowledge bases
- A pipeline approach enabling implementation and combination of modules for specific geoparsing applications
- GUIs for gazetteer browsing and testing developed modules

This makes HeidelPlace a unique and valuable tool for experimenting with new geoparsing approaches.

Generic Gazetteer Data Model

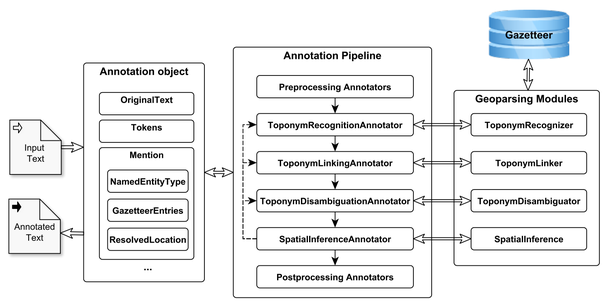


The Architecture of HeidelPlace

The architecture of HeidelPlace consists of three major components:

An Annotation Pipeline executes the geoparsing process on input documents. It utilizes the Stanford CoreNLP toolkit[1]. An annotation object contains metadata for a processed document. This enables passing information between annotators, which represent processing tasks that operate on a document. An annotation pipeline iteratively executes the tasks.

Geoparsing Modules unify the geoparsing process. For each step in the process, a module interface is defined that specifies expected in- and output and decides when the step can be executed.

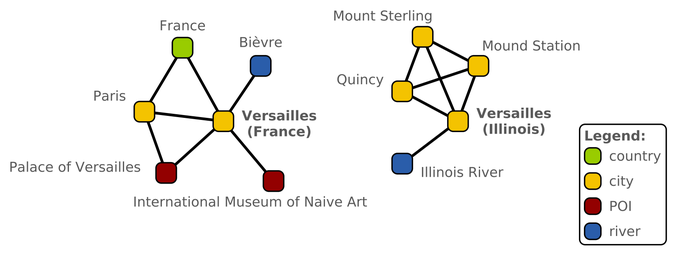


A Gazetteer serves as a knowledge base for the geoparsing modules. It employs a generic gazetteer data model, which provides a wide spectrum of place information. A flexible query system allows to efficiently search the gazetteer for places with certain characteristics.

This architecture forms a framework for geoparsing that can be easily extended and customized.

Outlook

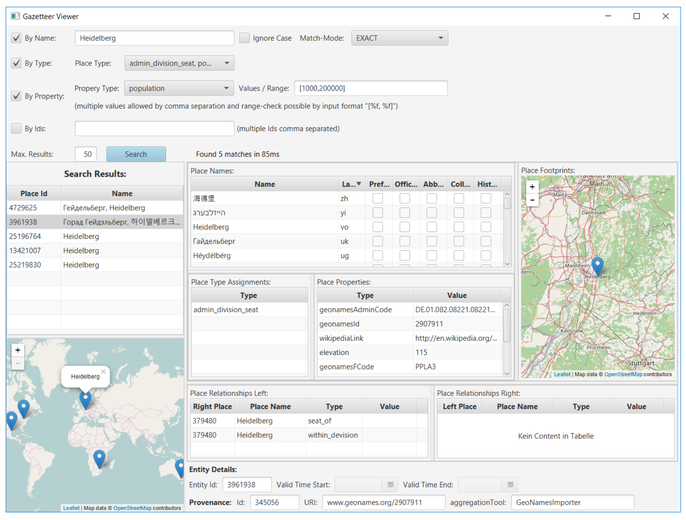
Getting ready for practical use:
Include more modules for state-of-the-art methods, e.g., co-occurrence based toponym disambiguation[2].



Further extensions of HeidelPlace:

- Quantitative evaluation framework
- Gazetteer web service
- UIMA component
- Gazetteer data editor

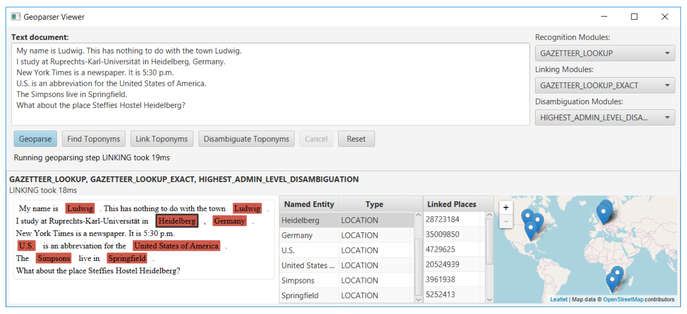
Gazetteer Browsing



Desktop App Gazetteer Viewer:

- Selection of search filters
- Detailed view of place information
- Intuitive navigation and visualization

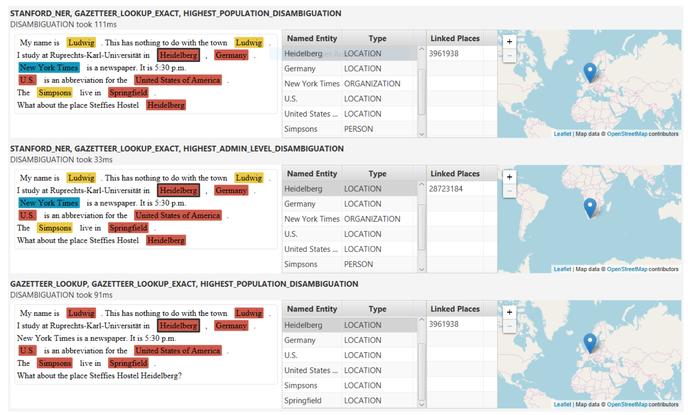
Performing & Analyzing Geoparsing



Desktop App Geoparser Viewer:

- Text input provided by user
- Selection of (pre-configured) modules
- Step-by-step geoparsing
- Output visualization for each step
- Interactive exploration of geoparsing

Comparing Geoparsing Methods



Desktop App Geoparser Viewer:

- Run multiple configurations at once
- Demonstrate handling of corner cases
- Understand differences of methods
- Experiment with module combinations

References

[1] C. D. Manning, et al.: **The Stanford CoreNLP Natural Language Processing Toolkit**. *ACL'14*, 2014

[2] A. Spitz, J. Geiß, and M. Gertz: **So Far Away and Yet so Close: Augmenting Toponym Disambiguation and Similarity With Text-Based Networks**. *GeoRich'16*, 2016

Contact Information:

Ludwig Richter

ludwig.richter@posteo.de

http://event.ifi.uni-heidelberg.de

This work was presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP '17), Sept. 7-11, 2017, Copenhagen, Denmark.