CLAVIN is an open source software application for document geotagging and geoparsing. It automatically extracts location names from unstructured text and resolves them against a gazetteer to produce data-rich geographic entities. It’s fast, accurate, and scales to accommodate big data in the cloud.

CLAVIN combines various open source tools with natural language processing techniques to extract and resolve geospatial entities from text, intelligently, accurately, and automatically. It handles misspellings, alternate names, and ambiguous references like “Springfield” or “Portland.”

By enriching documents with structured geo data, CLAVIN enables advanced geospatial analytics on massive volumes of unstructured text without massive cost implications.

To download the source code, or try out the interactive online demo, please visit: clavin.bericotechnologies.com
How CLAVIN Works
CLAVIN takes as input one or more completely unstructured text documents. A list of location names is first extracted from the text using a third-party named entity recognition tool, such as Apache OpenNLP Name Finder or Stanford NER. This list is passed to the CLAVIN location resolver, which then compares the location names against an Apache Lucene index built from the GeoNames.org worldwide gazetteer. CLAVIN selects the most appropriate matches for these place names based on the surrounding context of the document, and returns data-rich geographic entities representing the locations mentioned in the text. These entities contain a variety of data attributes (including lat/lon coordinates, country codes, administrative subdivisions, alternative names, population, elevation, etc.) that can be used to power hierarchical geospatial search and advanced geospatial analytics on unstructured data.

Advanced Features
CLAVIN utilizes a variety of natural language processing (NLP) techniques to derive geo data from unstructured text. Named entity recognition distinguishes “Grover Cleveland” from a city in Ohio, while fuzzy matching is used to capture misspelled location names, including phonetic spelling and typographical errors. CLAVIN recognizes alternate names for the same entity (e.g., “Ivory Coast” and “Côte d’Ivoire”), and intelligently disambiguates between ambiguous location names like “Springfield” based on the semantic context.

Using the GeoNames.org geographic database as its data source allows CLAVIN to have worldwide coverage spanning dozens of gazetteers consisting of well over 8 million locations and over 15 million unique location names.

Building CLAVIN from open source technologies and releasing it under the Apache License allows Berico to deliver these powerful and highly scalable capabilities with zero licensing cost.

In October 2012, CLAVIN was released to the public under the Apache License as Berico’s first official open source project. This means CLAVIN can be deployed to as many Hadoop nodes as required to fit your big data needs without having to worry about expensive enterprise licenses or costly usage fees.

We’d love to help you unlock the geospatial potential of your unstructured data by tailoring a custom solution based on CLAVIN. Contact us today at clavin@bericotechnologies.com.