

**Geosemble Technologies**

**White Paper on Organizing and Searching Documents by Geographic Location**

September 18, 2011

Version 1.1

Geosemble Technologies Inc.

2041 Rosecrans Ave. Suite 245

El Segundo CA 90245

310-414-9849 x400

[info@geosemble.com](mailto:info@geosemble.com)

**GeoXray**

As the number text documents continues to grow, the problem of organizing and searching documents becomes increasingly important. Search engines, such as Google, provide the ability to locate documents by keywords, but in many cases it is not obvious what keywords to use to find relevant information. Geosemble’s GeoXray product provides a scalable approach to organizing and locating text documents by location. In this white paper we describe the core capabilities underlying GeoXray and compare these capabilities to what is available in other systems for linking documents to locations.

**Linking Documents to Locations**

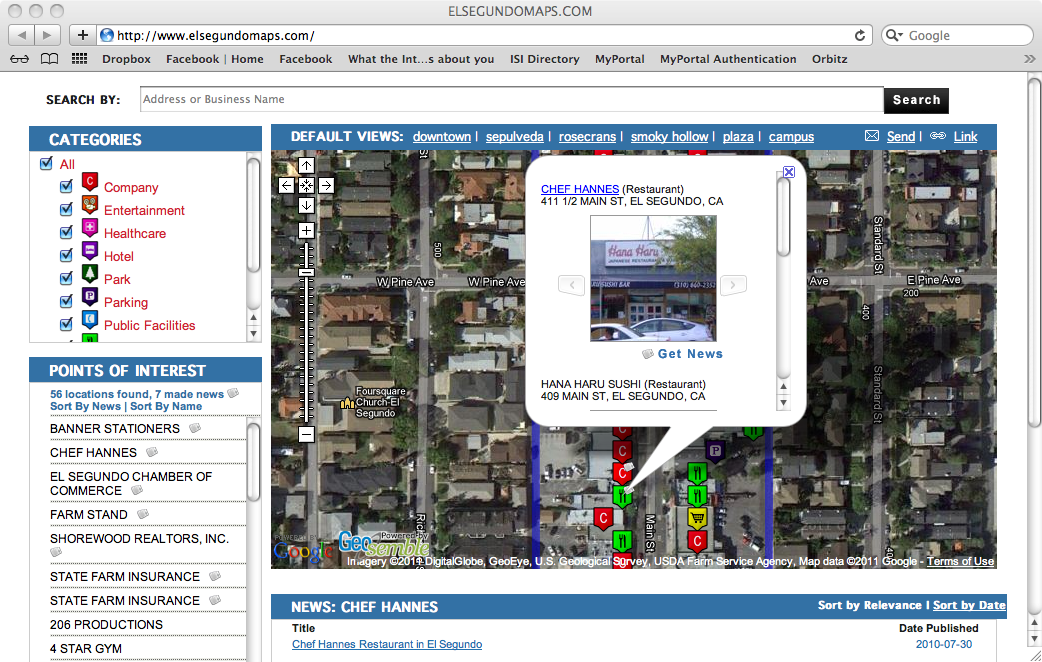
A core capability behind GeoXray is the ability to take a set of documents and link them to locations. GeoXray does this by considering all of the terms mentioned in a document. This can include not only the geographic terms, but also terms that mention related information, such as names of business, street names, and even phone numbers. The system considers the complete set of terms to then compute the most likely location that is the focus of the document. The primary geographic location of a document is referred to as the *geographic focus*. When GeoXray computes the geographic focus of a document, it also computes a corresponding score, which captures the confidence that the document is about a given location. In some cases a document may have more than one geographic focus and it would assign a score to each location. In general, just because a geographic location is mentioned in a document, it would not link it to that location. Rather, there would need to be sufficient evidence in the document that the location was a topic of the document.

A key difference of GeoXray with other commercial products, such as MetaCarta or Yahoo! Placemaker is that these other products do not actually determine the geographic focus of a document. Instead, these other systems take every geographic reference in a document and compute the corresponding geocoordinates of that geographic reference. In many cases, these references can be ambiguous and the linking based on a single reference can lead to incorrect results. Since these systems link every geographic reference, it means that for any given document one could have hundreds (or even thousands) of references linked to locations. Since these systems do not take a position on the geographic focus of a document, this means that even a passing reference to a location will result in a document being linked to that location. Then when a user requests all of the documents about a location, there will be many documents that are not actually about that location, but simply mention that location name in passing. And many of these documents will actually refer to other locations with the same or similar name.

**Fine-grained Linking**

Another core capability behind GeoXray is the ability to perform very fine-grained linking to locations. This means that instead of merely linking documents to a city or general area, GeoXray can link documents down to specific addresses or individual businesses. The system is able to perform this fine-grained linking by using a highly detailed gazetteer, which is a database of locations, for a region and then computing the documents that link to each of the individual locations. The fact that GeoXray uses the geographic focus for the documents makes this fine-grained linking possible. Otherwise, the system would link many irrelevant documents to each location. This ability to perform fine-grained linking makes it possible to build applications where linking down to specific buildings or addresses is required.

One example of such an application that requires fine-grained linking is an application that Geosemble built for the City of El Segundo, California for supporting business development for the city. In this application, current news about businesses in El Segundo is directly linked to each of the businesses. This application is shown below and can be viewed online at <http://www.elsegundomaps.com/>.



**Geographic Document Search**

The ability to determine the geographic focus of a document and perform fine-grained linking to locations directly supports one of the most important capabilities within GeoXray, which is the ability to perform geographic document search. In many cases, the required task is to find a set of documents about a specific topic within a specific geographic region. In some cases, that geographic region might be a specific building, such as a nuclear reactor, or in other cases the region could be a small area or even a city. GeoXray allows you to define the geographic region and specific set of non-geographic keywords. The system will then find all documents that mention those keywords that fall within the specified geographic region.

**Specialized Gazetteers**

The GeoXray product comes with a general gazetteer that covers the world. But in many cases, a user might be interested in a particular region of the world that requires a highly specialized gazetteer in order to obtain the needed accuracy in terms of document location. This might require a gazetteer in a foreign language or a gazetteer that contains highly-detailed knowledge about that region of the world. GeoXray can support both foreign language text in a gazetteer and can support highly-detailed user-provided data to construct specialized gazetteers for specific applications. If such a gazetteer is available, it can simply be plugged into the system and then applied to the relevant documents.

**Tools for Building Specialized Gazetteers**

In some cases where a specialized gazetteer is required, no such gazetteer will already exist. To address this problem, GeoXray comes with a set of tools for building gazetteers from existing sources. These may be databases already held by a customer or even from online sources that contain the required data. Geosemble has also developed tools for extracting data from maps, which contain a wealth of location data. By exploiting the available sources about a region, it is possibly to quickly build highly detailed gazetteers that will support accurate linking of documents within a given region

**Extracting and Linking Online Documents**

There is a wealth of data available online from news articles and web pages to blog posts and tweats. Geosemble has a set of tools available that support the real-time extraction of online data. Such tools make it possible to, for example, download the new articles for a particular region from local news providers on a daily or even hourly basis. These tools can be licensed and used directly by a customer or Geosemble can provide this capability as a service. Geosemble currently aggregates news articles from hundreds of Web sites on a daily basis. The extracted information can then be fed directly into GeoXray to support up-to-the-minute geographic document linking and search to support real-time intelligence gathering.