

Geographic text search improves surveyors' access to information

By John-Henry Gross, Product Manager, MetaCarta



A surveyor prepares to work at the proposed site for a new highway. He knows that part of the landscape is considered ecologically fragile—but what part? That information is contained in EPA documents. The Army Corps of Engineers was involved in the construction of a bridge nearby—what do they know about the area? Another part of the road runs by an abandoned mine. How will that affect this new highway project?

Surveyors need all this information to efficiently perform their work. But, the information may not be located conveniently in a database. Much of it is contained in unstructured documents generated by federal, state and local agencies, as well as by private sector companies and nonprofit organizations. An estimated 80 percent of all digital documents include at least one geospatial reference.

Unstructured documents within the enterprise and on the Internet are growing at an exponential rate of 50 to 60 percent per year. Even when relevant information resides within a structured database, the geographic references are often not available with the geographic coordinates necessary to query it geospatially. Only when both the structured information contained in databases and unstructured information are fused into a single cohesive view can a surveyor have a complete picture of a situation.

In the surveying and mapping fields, where activity is tied to a physical location, the geographic references in documents, reports, e-mails and blogs are important—but not necessarily easy to find. Most often, the data are collected by many different people, in different formats, both structured and unstructured. Few documents, if any, are clearly marked with a latitude or longitude.

When a surveyor performs an existing record search for documents describing restricted use or easements, searching through archives can be time consuming and, frustrating.



Traditional search engines are insufficient, as they cannot resolve ambiguous location references. Take, for example, the surveyor who is searching for a reference to Franklin Street located in a city named Lincoln. Traditional search tools cannot distinguish between the use of Franklin or Lincoln as names of places and as names of people. The results for such a search could easily number into the thousands, most of which would have nothing to do with the location in question. Equally difficult is a search for a Thomas Avenue, in a town whose former mayor's surname was Thomas. The sheer magnitude of the task can be overwhelming without the proper tools.

The tools do exist. The geographic text search (GTS) tool offered by MetaCarta is a geographically aware search tool that can disambiguate geographic terms and automatically resolve them to a specific latitude and longitude. GTS uses the geospatial information inherent in public works documents along with the non-geographic text, so officials can quickly determine whether a document refers to his or her area of responsibility and interest. In addition, the geographic pattern of the information becomes obvious

GIS perspectives

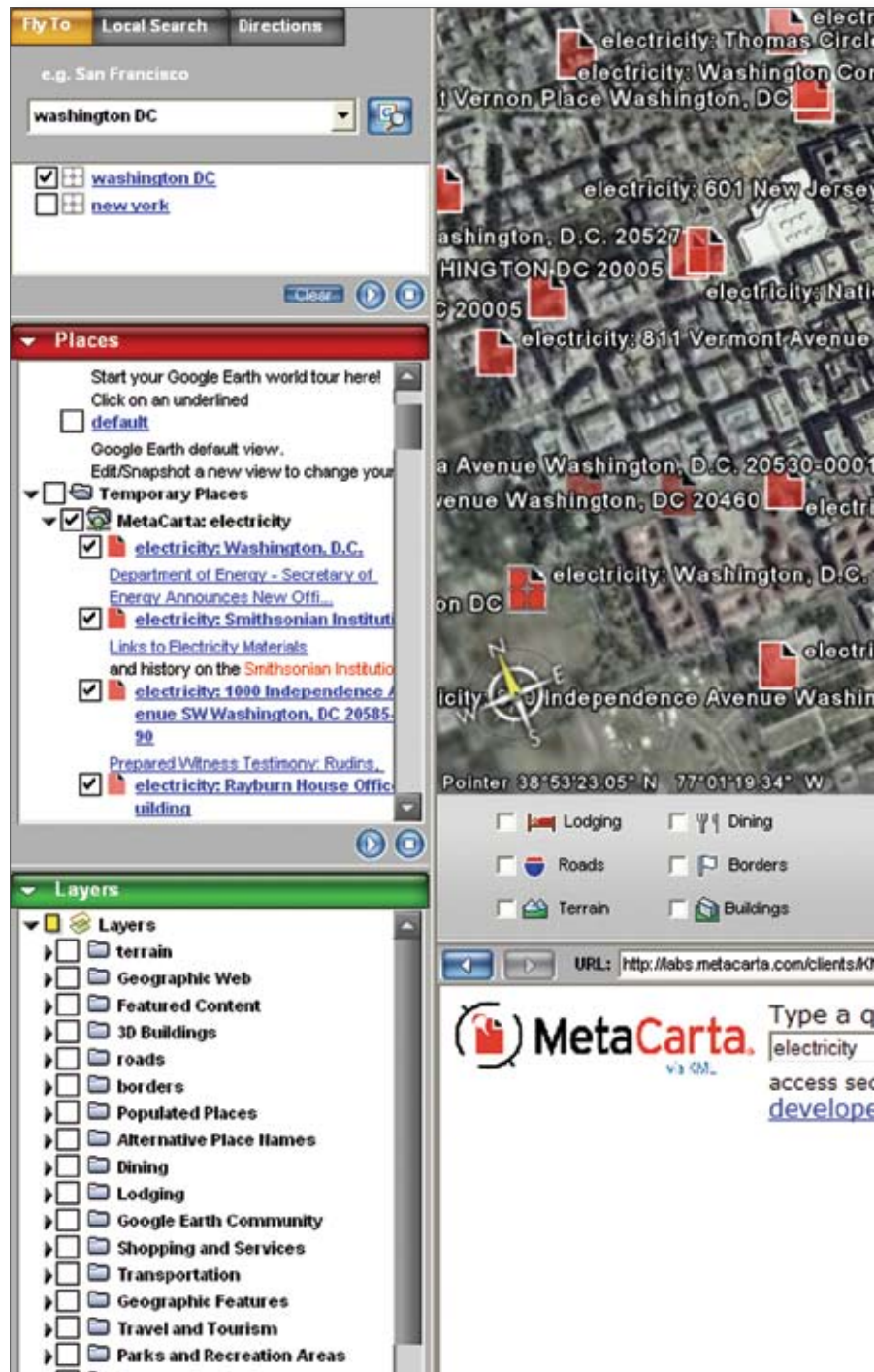
when the results are visualized on a map. MetaCarta GTS connects to GIS applications including ESRI's ArcMap, ObjectFX's SpatialFX, CRF's Terrain Analysis System (TAS), Google Earth and Intergraph's GeoMedia, and to content management systems such as EMC/Documentum and Open Text Livelink.

MetaCarta GTS is already in use at federal, state and local government agencies. The power of this solution comes from its ability to use natural language processing (NLP) to mimic the processes humans use to resolve an ambiguous location reference within a document and geospatially tag it with the correct latitude and longitude.

Human beings are easily able to distinguish ambiguous references. As an example, for a person in the Washington, D.C. metropolitan area, the statement "I went to Vienna last night," clearly refers to the Fairfax County, Virginia, suburb, not to the capital of Austria. But in fact, there are 35 separate places in the world named Vienna, with 31 of them in the continental United States. So, while a human can easily make the distinction between those references, a computer will have more difficulty.

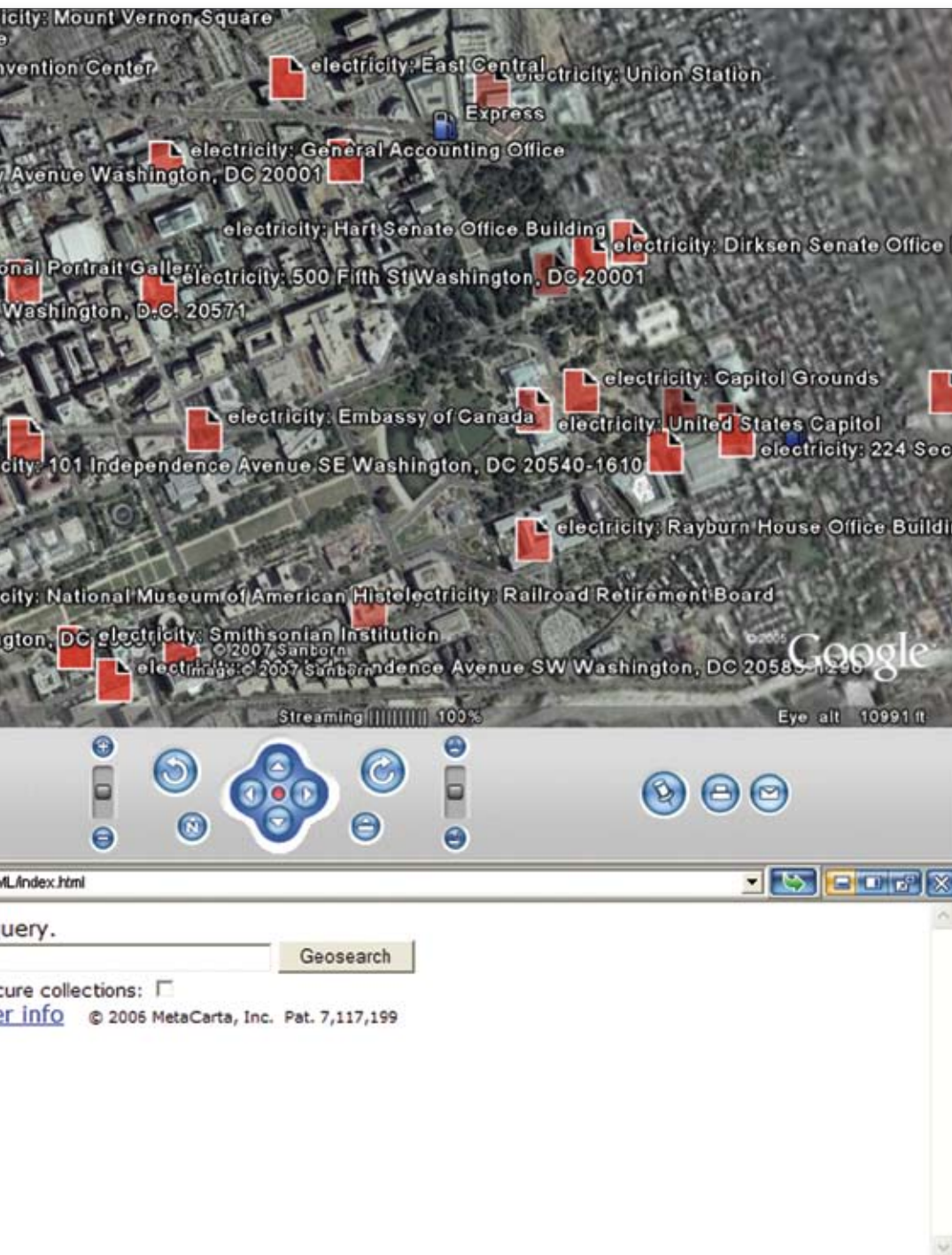
Using natural language processing, which is a subfield of artificial intelligence and linguistics, allows a computer to use context clues in a document to correctly determine the geographic area being referenced. For example, if the speaker continued the comment about Vienna by saying "then I drove through Tysons Corner to get home," the technology offers a significant degree of confidence that the speaker is indeed talking about Vienna, Virginia, as Tysons Corner is a neighboring suburb. However, if the sentence instead continued, "and we drove from there to Slovakia," there is a degree of confidence that the Vienna in question is in fact in Austria.

Also, natural language processing offers the ability to differentiate between references that may or may not be place names. The word "Denton" could refer to any one of 39 geographic locations—or it could be a person's name. Once again, the context of the surrounding document provides indicators. The words "city of" or "mayor of" preceding a name like Denton, or the words "community college" following, are strong positive indicators that



the candidate name is geographic. The words "Mr.", "Dr.", or a common first name preceding the name, are strong negative indicators.

By looking at a document's context, MetaCarta can distinguish between different places with the same name and people's names that are similar to place names. Furthermore, MetaCarta can identify relative references such as "30 miles south of Boise," tagging the most likely point of reference. The solution is thereby able to provide meaning resolution—discerning the actual intended meaning of the document's author.



Search results are literally viewed on a map, providing an intuitive picture of the geospatial pattern of those results. Results from a search query appear as icons on a digital map and in a results list. The location of each document icon coincides with the geographic locations mentioned within the document. If a user wants to find all documents relative to a geographic area—such as within a country, city, or latitude / longitude bounding box—the solution renders a map appropriately marked with icons representing every document that includes text pertaining to the identified location. By clicking on a document icon, a user gains direct access to the original document.

For surveyors, the benefit of fusing geospatial information and text search is clear. Before embarking on a project, the surveyor can first gather all of the necessary information from disparate government and private sector sources through one or two searches, rather than laboriously searching through the archives of every organization that may have the information in question. Furthermore, the surveyor can target his or her search to the specific city, neighborhood, or

even street of interest, eliminating information from locations that are not relevant to the current project.

With the geographic references tagged in the document, the search can then be controlled to return only the documents containing references to the geographic area of interest. Users can select a country, state, or city to search, or they can further limit the search with artificial boundaries to select a region or neighborhood. Once the geographic area is defined, the user enters search terms. The solution returns only documents with those search terms that are tagged with latitudes and longitudes within the geographic area of interest.

even street of interest, eliminating information from locations that are not relevant to the current project.

To produce the highest quality, most accurate maps, a surveyor must have all pertinent information available in both structured and unstructured documents. The use of a geographic text search solution to intelligently search unstructured and structured data containing geographic terms ensures that surveyors can effectively, efficiently, and accurately gather all the information needed to complete a project. MetaCarta's GTS is that tool.