Title: Mapping Ideas from Cyberspace to Realspace with Geospatial Fingerprints.

Ming-Hsiang Tsou, mtsou@mail.sdsu.edu, Department of Geography, San Diego State University, U.S.A.
(Principle Investigator of NSF CDI-Type II Award # 1028177).

With the new media (e.g., the Web and online social networks), scientists can trace geographic and chronological patterns to reveal the nature of significant events such as radical concepts or epidemic outbreaks. The Web, a powerful platform for collective thinking and idea exchange, provides valuable intelligence to help scientists monitoring processes ranging in diversity from the spread of diseases to the structure of terrorist networks. The information dynamics can be transformed into visual maps and information landscapes for space-time analysis. Understanding the diffusion and cluster patterns in response to terrorist movements and epidemics may significantly facilitate intervention, and eventually, prevention responses. With the support of the new NSF-CDI funding (from 2010 – 2014), our research team developed a Spatial Web Automatic Reasoning and Mapping System (SWARMS) prototype for visualizing and analyzing the spread of concepts, ideas, and events in cyberspace (http://mappingideas.sdsu.edu). Hundreds of web pages were geocoded with real world coordinates and represented in the form of web information landscapes. These web information landscapes (maps) can help us monitor spatial and temporal patterns that coincide to reveal the nature of significant events, radical concepts or epidemics. These dynamic web information landscapes can reveal unique "geospatial fingerprints" associated with the contexts of different keywords.

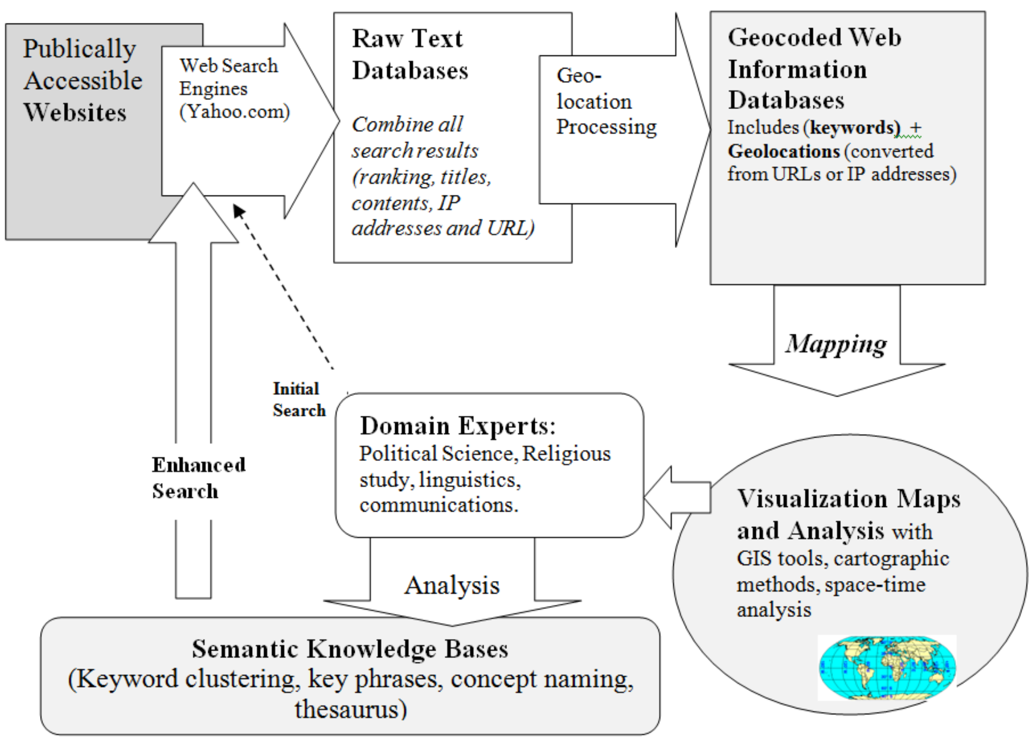

Figure 1. The Spatial Web Automatic Reasoning and Mapping System (SWARMS) framework.

Figure 1 illustrates the overall SWARM prototype framework. First, we performed initial searches using pre-defined keywords on specific topics (e.g. infectious diseases or radical concepts) provided by domain experts to search for the top 1000 hits from publically accessible websites (using the Yahoo search engine). Then we converted the search results into a *[Raw Text Database],* which included all web search results (ranking, titles, IP addresses, and URLs). The system then uses the IP addresses to convert raw texts into *[Geocoded Web Information Databases],* adding geospatial locations (latitudes and longitudes) for each record. By utilizing WHOIS protocol and GIS software (ESRI ArcGIS), we converted the geocoded databases (created by a Microsoft SQL server) to *[Visualization maps]* showing the information landscapes of specific ideas or keywords. We then applied advanced GIS analysis and visualization methods to understand the dynamic change of these concepts and events over space and time. Computational linguistics experts can review the resulting maps and then establish frequencies of occurrences of “key terms,” separately and in clusters. Multiple *[Semantic Knowledge bases]* related to ideas, concepts and special topics can be created and revised based on the visualization maps and further space-time analysis. The revised keyword clusters and phrases will be used for the next round of Web query process. New web pages and websites will be discovered in this iterative process by refining keyword clusters and analyzing new information landscapes (Figure 1).

Our research team used GIS kernel point density functions to illustrate "hotspots" and cluster patterns of related web pages with their keyword search ranking numbers. ArcGIS software package were used for the kernel density calculation by specifying map unit threshold (radius) and map unit output scale in the maps. The resulting maps can show the popularity of the queried keywords among different cities and regions. In our test, we found that a single web information landscape may only provide limited information about the context of selected keyword. Comparing multiple web information landscapes with standardized kernel density methods might reveal important "geospatial fingerprints" corresponding to the context of selected keywords and concepts. In our test, two selected keywords, "Jerry Sanders" and "Antonio Villaraigosa", have implicit spatial relationships with the City of San Diego and the City of Los Angeles (Jerry is the mayor of San Diego and Antonio is the mayor of L.A). These implicit geospatial relationships can be visualized in their geospatial fingerprints when comparing the differences between the two web information landscapes. In our research, geospatial fingerprints are defined as the unique spatial patterns and clusters of information landscapes associated with different keywords and concepts.

We also realize that our mapping methods are still very premature and need to be improved. For example, we need to exclude several search hits from Wikipedia websites and public news websites. Also, the geolocation method in the SWARMS can only convert 90% of URLs into latitude and longitude coordinates. The rest of 10% of websites were not used in the visualization maps yet. Moreover, our current SWARMS prototype can only use Yahoo and Bing search engines because the Google Search Engine limits the API search up to 64 records only. Hopefully, these limits and restrictions will be addressed in our future research. During our testing and prototyping, we also find out that the following three research topics are essential to the future development of cyberspace mapping.

1. How to analyze the spatial relationships among **points** (websites and individual web pages), **networks** (hyperlinks within web pages), and **regions** (community groups or social network groups) on cyberspace?
2. How to develop effective cartographic representation methods and map symbols to illustrate the dynamic flows of ideas and concepts on the Internet?
3. Which spatial scale is more appropriate for representing specific ideas or concepts on maps?

These research questions along with the new SWARMS research framework may help us explore the dynamic interactions of various ideas in cyberspace and develop more advanced visualization principles and methods for cyberspace mapping.

ACKNOWLEDGEMENT

The SWARMS framework is developed by the funding support from NSF project “*CDI-Type II: Mapping Cyberspace to Realspace: Visualizing and Understanding the Spatiotemporal Dynamics of Global Diffusion of Ideas and the Semantic Web*”. (NSF Award#_1028177). This project is a multidisciplinary collaboration among Ming-Hsiang Tsou (Geography), Dipak Gupta (Political Science), Jean Marc Gawron (Linguistic), Brian Spitzberg (Communication) and Li An (Geography) at San Diego State University.