SEMI-AUTOMATED VISUALIZATION OF SPATIAL INFORMATION IN UNSTRUCTURED TEXT

by

Sarah Marie Gehring

May 2015

A Thesis Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Special thanks to Joanna for her numerous editing sessions she squeezed in while caring for a newborn at home. Without her input and encouragement, I do not know if I would have ever completed this!

Thank you to my committee members Jennifer Swift and Bob Vos for your contributions to the content of the paper, these suggestions helped tie it all together. Thanks to my writing instructor Mariko for helping teaching me some grammar and writing mechanics I had forgotten and helping to polish my writing.

Thanks to my family and friends for listening to me talk of this elusive thesis for over a year before I really started to work on it. Their support and advice helped pull me through when I felt like giving up.

Most of all, thank you to my advisor Yao-Yi for putting up with my terrible attitude and lack of motivation for the first full-year before something finally clicked and I was able to actually produce some quality work. I know for a fact that I was a challenge to work with but somehow he put up with me. With his guidance, I was able to complete this project.

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES: APPENDIX A

LIST OF FIGURES: APPENDIX B

LIST OF ABBREVIATIONS

CLAVIN	Cartographic Location And Vicinity INdexer
CSV	Comma Separated Values
DIO	Data Integration and Output tool
DIP	Data Integration and Processing
GIS	Geographic Information System(s)
MIT	Massachusetts Intitule of Technology
NER	Named Entity Recognition
NLP	Natural Language Processing
OSM	OpenStreeMap
PDF	Portable Document Format
PIG	Parse, Identify & Geo-lookup
SAV	Semi-Automated Visualization
STEWARD	Spatio-Textual Extraction on the Web Aiding Retrieval of Documents
TXT	File extension for a Text File
URL	Uniform Resource Locator
WGS84	World Geodetic System 1984

ABSTRACT

Digital information with a spatial component is being generated at an astounding rate, from sources such as Flickr Videos, online news, and “tweets” on Twitter. The ability to identify locations in unstructured text and quickly generate a map unlocks valuable information about the context of the locations in the text. Geoparsing, the process of assigning geographic coordinates or other geographic identifiers to unstructured text, extracts this valuable information from text (Nikolajevs and Jekabsons 2013). Existing studies and tools focus on the challenges of location extraction and disambiguation. These studies do not focus on visualizing the extracted locations, and generally use a simple method of displaying each location as a single point on a map. This thesis examines the current geoparsing text-to-map applications, identifies challenges to generate a map from a text document, and defines an approach to display locations with boundaries and relationships between locations on a map. The outcome of this thesis is a semi-automated geoparsing, data integration, and visualization application to convert the locations in text-based news articles to locations on a map. This approach provides an efficient and effective way to display the spatial context of a text document and allow for interpretations of the data that is not readily apparent from the text by itself.

CHAPTER ONE: INTRODUCTION

Location references in text are seen everywhere; they are evident in social media, such as a person sharing that he is are having dinner at his favorite restaurant in San Francisco. The writer uses the reference to orient the reader to a specific location where a story occurs. In this situation, San Francisco is a geographic place with a boundary surrounding the city that, when shown on a map, provides the viewer with geographic context to the location. When there are several locations in a text document, displaying them on a map can help the viewer understand the relationship between locations. For instance, San Francisco is within California, and is adjacent to Oakland. To display these locations on a map, there are a wide variety of geographic techniques available.

According to Dodge et al., geographic visualization is a way to represent information about an area that cannot be seen directly (Dodge, McDerby, and Turner 2008). Until recently, paper maps have been the primary form of geographic visualization and have been used as a way to organize and communicate spatial knowledge and provide a navigational guidance (Dodge, McDerby, and Turner 2008). Since the 1960s, cartography and visualization techniques have evolved. Computers and Geographic Information Systems (GIS) have become increasingly important tools in geographic visualization by providing many dynamic mapping solutions (McMaster 2005). This systematic approach to mapping provides endless opportunities to identify, store, and process spatial data.

Geoparsing is a process for assigning geographic coordinates to words or phrases in text that requires first identifying the location text (e.g., place names), then second assigning the geographic coordinates to the text (Nikolajevs and Jekabsons 2013). Using the following sentence as an example: “A man was arrested in Concord, a suburb of San Francisco, for

allegedly shooting another man," geoparsing identifies Concord and San Francisco as locations. Geoparsing then assigns these cities their corresponding longitude and latitude coordinates and plots them on a basemap. GIS and geoparsing enable information visualization, the visual representation and analysis of nonnumeric abstract information (McMaster 2005). Geoparsing transforms text documents that were once just words on a page into complex visual representations of all the locations presented on a map.

Some existing geoparsing tools focus on the extraction of locations from web pages to aid spatial search results, by using the location references to select search results that are spatially relevant to the search area such as Web-a-Where (Amitay et al. 2004) and Spatio-Textual Extraction on the Web Aiding Retrieval of Documents (STEWARD) (Lieberman et al. 2007). These tools that use geoparsing for spatial search do not focus on the visualization of the locations and may just show a point on the map to represent the search results or just use that spatial information for processing the search and not display it on a map. Chapter 2 will describe several geoparsing tools.

Tools like Yahoo! PlaceSpotter (2014c) and OpenCALAIS (2014b) and Cartographic Location And Vicinity INdexer (CLAVIN) (Berico Technologies 2014) perform geoparsing and have the capability to map each location as a point on a map (D'Ignazio et al. 2014). These tools provide geoparsing capabilities that utilize data from a gazetteer, a dictionary of geographic locations, to assign geographic coordinates to a location and plot each location as a point on a map. This method of geographic visualization provides some basic information to the viewer about the spatial proximity of the locations in text, but plots every location the same way, as simply a point on the map (Figure 1). Having only the gazetteer as a data source limits the resulting output to points.

Maps that plot locations as points command the user to extrapolate the spatial context and geographic relationships between various locations on their own. For example, in Figure 1 the user can extrapolate that two of the locations are relatively close to each other, near San Francisco, and the third location is near Riverside. The spatial extent of the map does not allow the viewer to see understand what other landmarks are nearby. Without labels or a legend to identify the points, there is no easy way to know the name of the locations. The next example provides a different approach to producing a map from text.

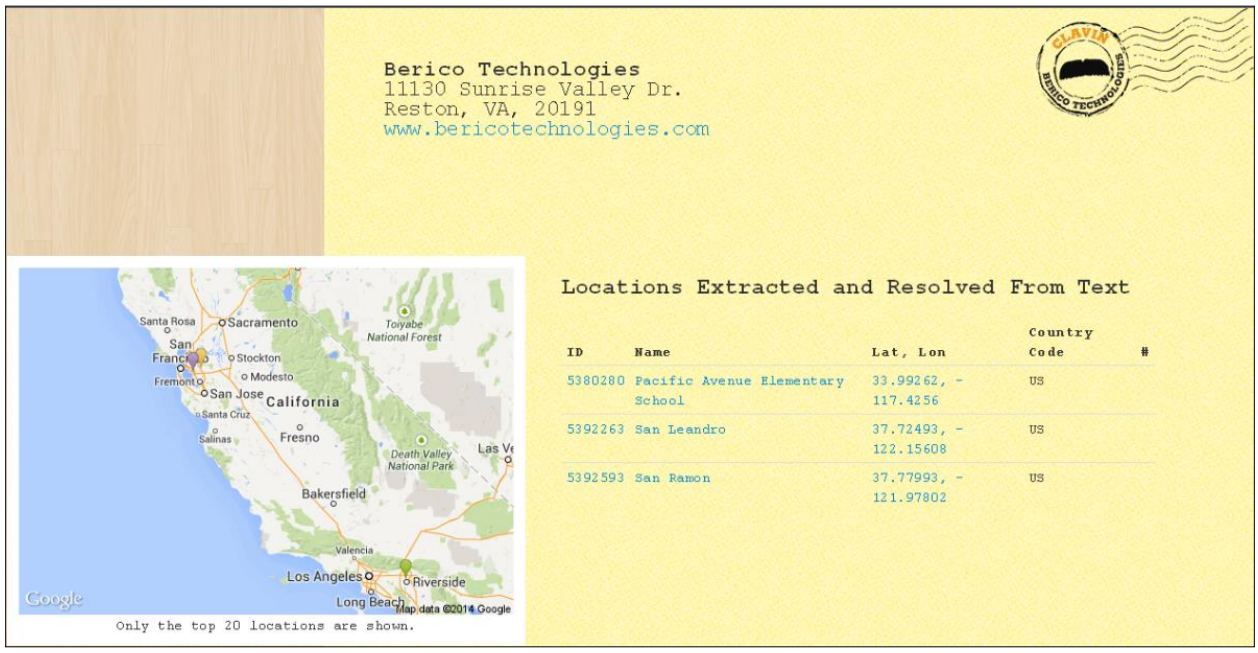


Figure 1 CLAVIN Online Tool Output
Image by Berico Technologies (Berico Technologies 2014)

This thesis presents an approach that integrates gazetteer data with additional data sources and exploits rules for how the integrated data should be displayed in a map to provide a complete picture of the spatial context and relationships between locations in text. For example, in Figure 2 below, San Ramon is within Contra Costa County and Walnut Creek is adjacent to

Concord. Current text-to-map extraction techniques using geoparsing with only gazetteer data are useful when identifying locations in text.

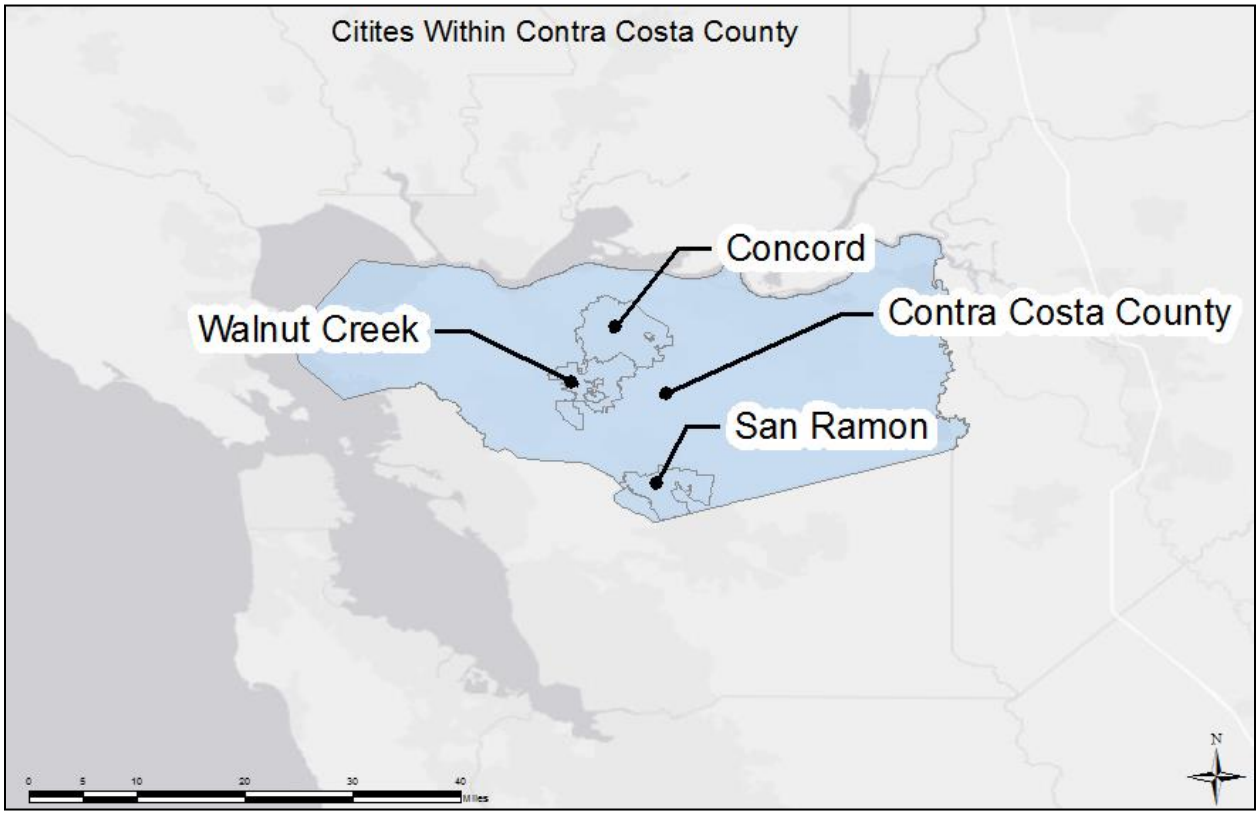


Figure 2 Map Example Showing the Relationship Between Cities and County

1.1. Problem Definition

To the knowledge of the author, there are no robust and complete solutions to automatically extract spatial information from unstructured text and present it on a map displaying spatial relationships between locations. The challenges to achieving this are: 1. The spatial entities in a document need to be extracted; 2. The locations need to be assigned spatial reference; 3. The spatial relationships between these entities need to be identified; and 4. The entities and their spatial relationships (e.g. a city in a county) must be rendered for visualization. Existing software solutions (described in detail in Chapter 2), which sufficiently solve the first problem of extracting the spatial entities from the text, but are inadequate at addressing

challenges two, three, and four. Considering the challenges above, there is currently a gap in text-to-map software. This thesis project aims to fill that gap by integrating a data source with location boundaries and creating rules and templates for how to display the data on a map.

1.2. Objectives

The goal of this thesis is to develop a method and a working prototype of a semi-automated application to: 1. geoparse text to identify locations, 2. integrate location boundary feature data with the identified locations, 3. identify relationships between locations, and 4. display the spatial information on a map. This thesis project focuses on a rule-based method to integrate the boundary data with the point data and plot the spatial information on a map.

This thesis project uses existing open source software to perform the location identification, location extraction, and association of geographic coordinates to locations in text documents (Section 3.4.1). The focus is the creation of a toolset to integrate the location boundary data with the point locations and the presentation of this information on a map. (Section 3.4.2). The final output are maps which display the boundaries of the locations mentioned in the text instead of showing points to represent them.

1.3. Motivation

The motivation for this research comes from an initial idea that it would be interesting to incorporate maps into e-books. This would be especially useful for books that have a strong spatial component to them and mention many locations. For example, the book *Wild: From Lost to Found on the Pacific Crest Trail* by Cheryl Strayed, is a memoir about a woman who hikes the the length of the U.S. West Coast. Since the story is about hiking along a trail, there are many spatial references along the way. In the paper and e-book versions, there is a one page map that shows the entire trail from the north end of Washington down to the south end of California

(Strayed 2013). This map provides an overview but does not allow the reader to discover much about the specific locations in the text as they read. When reading an e-book, such as Wild, it would be valuable for the reader to have the option to view the locations in the story on an interactive map where he could zoom in and out to understand the context of the location and see what other landmarks are nearby. The manual creation of maps for e-books would be a time consuming process; the use of automation to identify and map locations from text is a likely solution.

1.4. Scope

This thesis project uses online news articles as the text input to demonstrate the overall approach. To geoparse the articles, a custom version of an open-source software package is used since there are already existing solutions that do this portion well. The thesis focus is to create a toolset that integrates additional data sources and displays the spatial relationships between the locations on a map.

This chapter described the opportunity, motivation, and scope of this thesis project. The remainder of this thesis is organized into four additional chapters. Chapter 2 is an overview and literature review of the existing software and techniques that are available for location extraction from text. Descriptions of the methods and data sources are in Chapter 3, including tool customization and creation. Chapter 4 presents an experiment and evaluation of the results. Finally, Chapter 5 concludes the paper with a summary of the thesis project and a discussion of future work.

CHAPTER TWO: BACKGROUND AND LITERATURE REVIEW

This chapter looks at the current work to address the challenges, described in Section 1.1, of providing an effective text-to-map visualization and the choice of technologies for this thesis project. Section 2.1 covers geoparsing with subsections for: geoparsing gazetteers, parsing ambiguity, and available geoparsing software. Section 2.2 describes the choice of technology for this thesis project based on the review of existing technologies mentioned in this chapter.

Before the discussion of geoparsing, a brief overview of the disciplines of Natural Language Processing (NLP) and Named Entity Recognition (NER) is necessary. NLP is a discipline that studies how to use computational techniques to analyze human language and communicate with humans using their natural language (Cambria and White 2014). This is necessary for a computer system to be able to recognize locations. Cambria and White acknowledge that even after decades of research in NLP, machines still lack the ability to achieve the understanding of natural language that humans are able to derive (Rauch, Bukatin, and Baker 2003, Cambria and White 2014). Humans are able to use their past experience to understand the context of language and derive meaning from ambiguous text (Rauch, Bukatin, and Baker 2003).

NER is a sub-discipline within NLP with over 20 years of research. Recognizing named entities is not specific to geographic places. Early definitions of what constitutes a named entity limits it to proper names of persons, locations, or organizations (Nadeau and Sekine 2007). This definition has since grown to incorporate biological species, and temporal expressions (Nadeau and Sekine 2007). The next section describes geoparsing, an NER technique that focuses on geographic locations.

2.1. Geoparsing

Geoparsing is the technique of "recognizing geographic content" that is used to identify spatial information in unstructured text, including news articles (McCurley 2001). This subsection discusses gazetteer for geoparsing, parsing ambiguity of geoparsing results, and available geoparsing tools.

2.1.1. Gazetteer for Geoparsing

A gazetteer is a geographic dictionary that is the source of geographic named entities that NER recognizes. According to Amitay et al. (Amitay et al. 2004) a good gazetteer should not only contain a comprehensive listing of the locations that are relevant, but it should also contain alternate names and abbreviations so that the software is able to recognize aliases for the same location. For example, a good gazetteer understands that New York has several aliases such as NYC, and N.Y.C., and will recognize them as the same location.

It is easy to (incorrectly) assume that the more locations appearing in the gazetteer, the more likely a program is accurate in its identification. As Keller et al. point out, a gazetteer needs to be carefully designed by people with contextual knowledge; simply adding more terms to a gazetteer will likely result in more false positive results (Keller, Freifeld, and Brownstein 2009). Including more locations in a gazetteer will also affect performance and likely add time in the program's location identification process (Lieberman et al. 2007). Increasing the number of locations in a gazetteer may also increase the chances of locations becoming ambiguous because there are more locations with the same name.

2.1.2. Parsing Ambiguity of Geoparsing Results

When parsing data, extraction ambiguity comes in two forms: context ambiguity and location ambiguity. Context ambiguity occurs when a named entity is both a location and has another meaning, like a person's name. For example, NER software should be able to identify which of the many named entities called "Washington" is being referred to in the context of the text. The context of the sentence should provide enough information to be able to determine if a word such as "Washington" refers to a person like George Washington, or if it refers to a location, like the state of Washington.

Once geoparsing determines that "Washington" refers to a location in the current context and not a person such as George Washington, it is still possible that there is location ambiguity between more than one "Washington" location. In this example of location ambiguity, "Washington" could refer to the US state of Washington, Washington D.C., or one of many other "Washingtons".

To resolve location ambiguity Web-a-Where, a webpage geotagging software, represents the relationships between locations in the gazetteer in a hierarchy where there is a connection between different levels such as city, state, and county (Amitay et al. 2004). For example, Springfield is a city name with many different locations throughout the world. If Springfield appears in the text, and later in the text, Illinois appears preceding Springfield, then Illinois receives a higher confidence score based on the hierarchy relationship between Springfield and Illinois, rather than connecting Springfield to Connecticut. This is known as containment disambiguation because the location ambiguity is resolved by finding that one location (the state) contains another (the city) (Lieberman et al. 2007). This Springfield example uses spatial relationships to resolve location ambiguity.

Web-a-Where also reduces location ambiguity by limiting the size of the gazetteer to 40,000 entries, compared to other gazetteers that contain 1 or 2 million entries (Amitay et al. 2004). Limiting the number of locations reduces the number of ambiguous matches because this reduces the number of places with the same name. This may lead to inaccurate results because the dataset could be missing location information relevant to the region due to its small subset of locations.

According to Lieberman et al. another commonly used technique to resolve location ambiguity is to choose the location with the highest population (Lieberman, Samet, and Sankaranayananan 2010). This theory is operating on the principle that if a location has a large population, it is more likely to appear in the news and be the referenced location. A city of the same name may exist, but it has a smaller population and may only be relevant in local news. While useful on a macro-level, this would not be the best technique for identifying locations in local and regional news. Assigning geographic coordinates to locations is possible once the system resolves parsing ambiguity.

Geo-Lookup is a term used by Kelm et al. to describe the process of assigning geographic coordinates to a location's name (Kelm, Schmiedeke, and Sikora 2011). After a system identifies a location and resolves parsing ambiguity, each location can be given a geographic representation by assigning longitude and latitude coordinates by looking up the location in a gazetteer.

2.1.3. Available Geoparsing Software Packages

Gelernter et al. use a geoparsing machine-learning technique to identify locations in "tweets," in both English and Spanish, based on language rules (Gelernter and Zhang 2013).

Their approach shows geographic names can survive an English to Spanish translation and still be recognized by a Spanish Geoparser. Kelm et. al use geoparsing to extract locations from the meta-data of videos on Flickr and lookup their geographic coordinates in the GeoNames gazetteer as part of a three step approach to identify where a video was filmed (Kelm, Schmiedeke, and Sikora 2011). Their study's highlights include geoparsing the meta-data that provided the best results of the three techniques in their framework for mapping the filming location of Flickr videos.

There are several web services that provide geoparsing functionality. The output of geographic coordinates easily converts to points to plot on a map. Yahoo! PlaceSpotter is a fee-based application that provides web services developers can integrate into their applications to identify locations in text and return the centroid, as a point, of named places (2014c). OpenCalAIS is a robust web service that is not specific to locations; it can extract many types of named entities including locations (2014b).

CLAVIN is an example of a software program that uses NER to identify and extract locations from text, resolves each geographic location to the point found in the gazetteer, and plots each point on a map (Berico Technologies 2014). The CLAVIN online tool uses the foundation of Google Maps to plot each location as a point on map, as seen in Figure 1 above (Chapter 1). The location names and coordinates are next to the map in a list. If the user has no geographic knowledge of the area, then it would be difficult to distinguish between the points on the map because there are no labels. The CLAVIN map output can be improved by labeling the points on the map, finding the boundaries of the locations, and adjusting the scale of the map to provide additional details.

Another web-tool that utilizes geoparsing and the Google Maps mapping engine is STEWARD (Lieberman et al. 2007). STEWARD is a search engine, like Google, where a user types in a query and receives a list of results. STEWARD also returns a map with the list of results. The point to represent a result on the map is determined by finding the spatial focus of the web document. The spatial focus is determined by geoparsing the document and running an algorithm to decide which location reference is the focus of the article (Lieberman et al. 2007). Figure 3 shows an example of the results STEWARD provides, which displays the search results and ranking according to the spatial relevance STEWARD assigns.

Current studies of location extraction from news articles typically use geoparsing to identify all the locations in the articles to aid spatial search results and to retrieve articles relevant to the locations being searched. Existing software that extracts locations and plots the points on a map uses the basic technique of one per location. There are no known studies that focus on the relationships between identified locations and displaying the boundaries of locations. Chapter 3 will define the proposed method for the solution to fill this gap in the current text-to-map software.

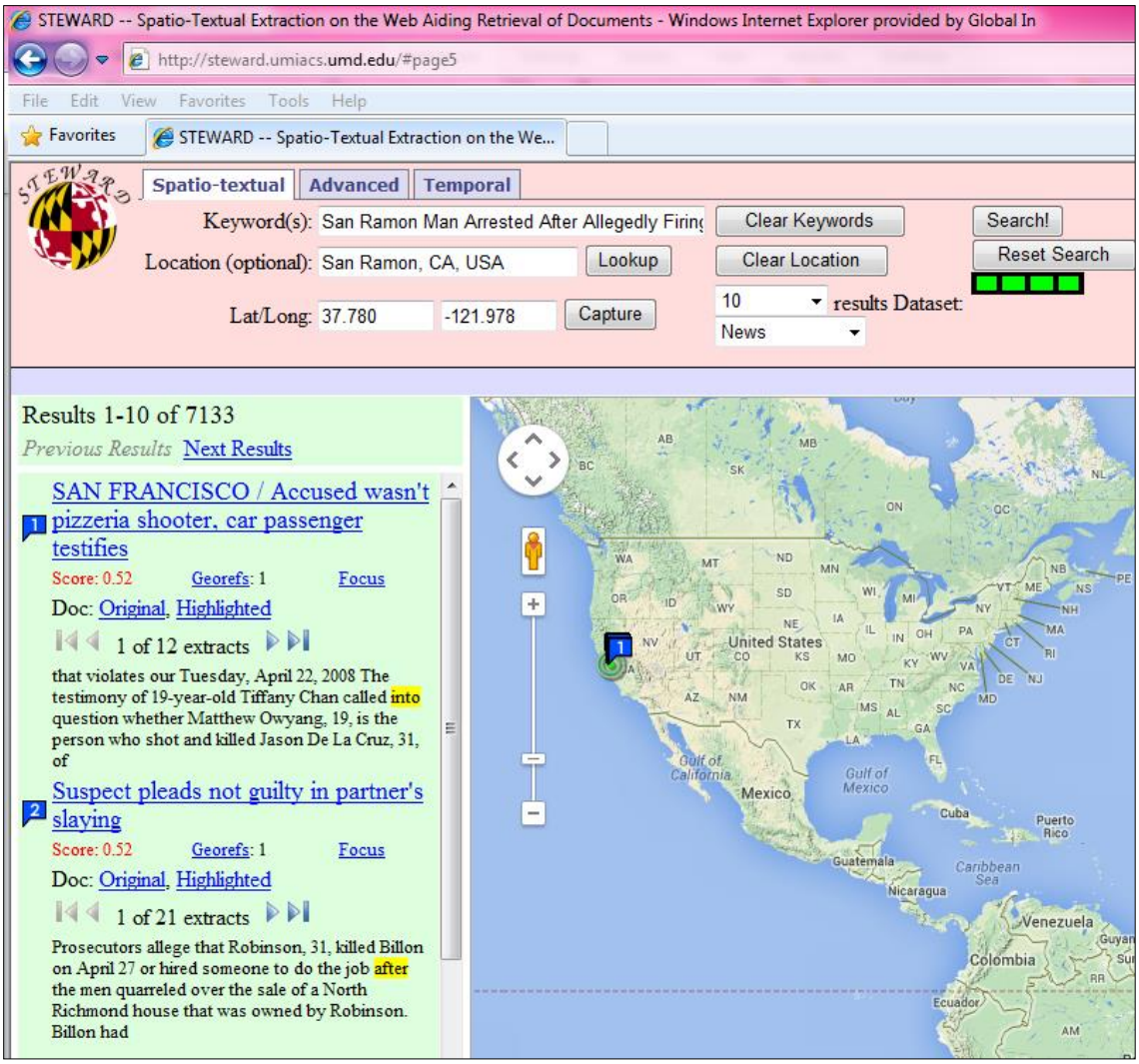


Figure 3 STEWARD Mapped Search Result of Webpages
Image by STEWARD

2.2. Choice of Technology

Berioco Technologies' CLAVIN was selected as the geoparsing and geo-lookup software because it provides a Java library for free download, which means programs can be written to utilize the geoparsing and geo-looking functionalities. There is also an online version of CLAVIN that provides a baseline to compare the map results of this thesis against. A study at The Massachusetts Institute of Technology (MIT) showed that CLAVIN performed as well as

the fee-based program, Yahoo! PlaceSpotter, and better than openCALAIS in terms of geoparsing results (D'Ignazio et al. 2014).

To integrate data and generate maps, ArcGIS version 10.2 (Esri 2014a) was chosen. Although it is a license-based software with a fee, ArcGIS Model Builder provides a graphical modeling tool which makes it easier to use than writing python code for every process. The model-based structure clearly shows the input and output variables and the flow of data through the process. For the purpose of this thesis project, a personal computer running 64 bit Windows 7 with eight Giga-bytes of RAM and both CLAVIN and ArcGIS10.2 installed was utilized.

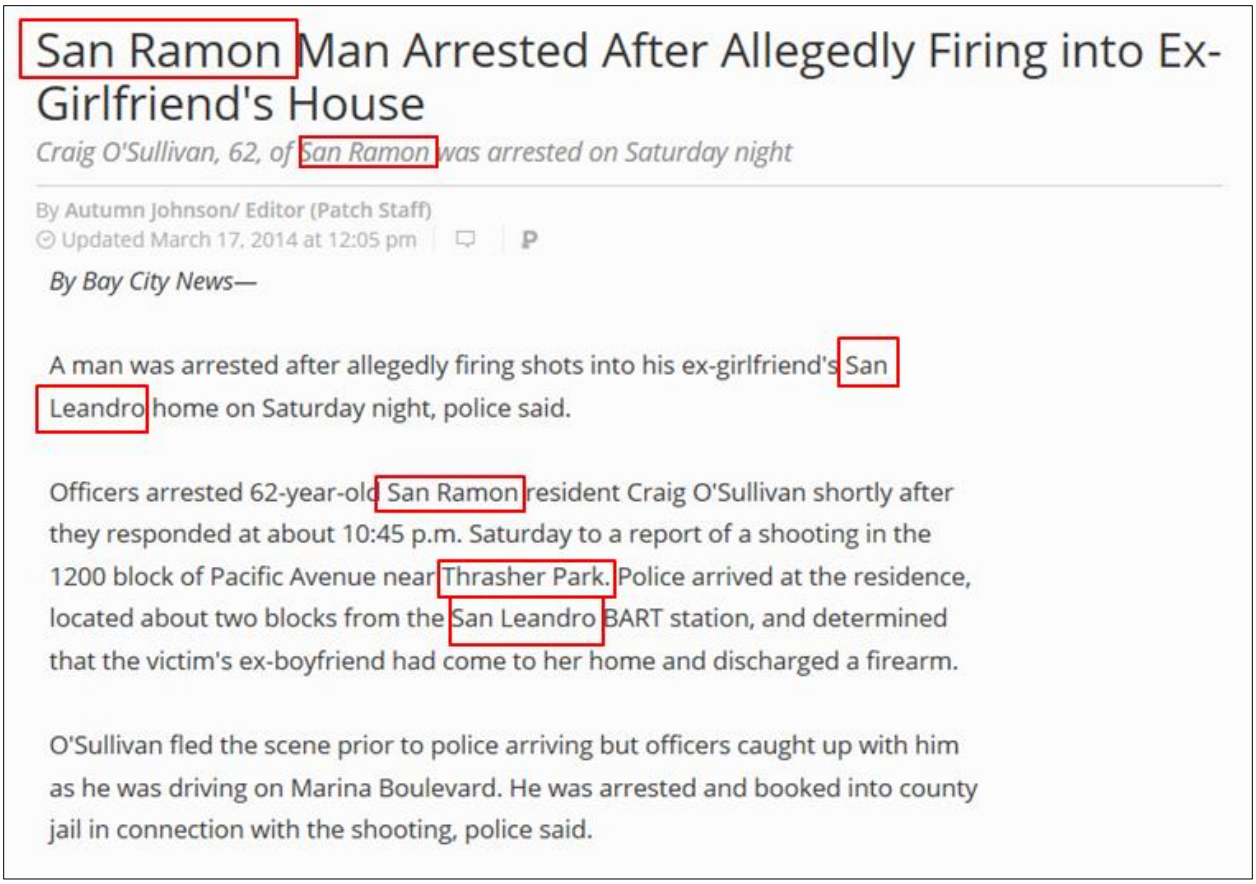
CHAPTER THREE: METHODS AND DATA SOURCES

This chapter presents the methods, data sources, and a case study to demonstrate the use of the Semi-Automated Visualization (SAV) application developed as part of this thesis project. The purpose of SAV is to extract locations from online news articles and display the boundaries of the locations on a map to provide spatial context and display relationships between locations.

3.1. Method Overview

The case study this chapter presents uses one article to demonstrate SAV's approach. Figure 4 shows the article used in the case study about the arrest of a man involved in a shooting. The red boxes identify the locations that geoparsing software recognizes and are not part of the original article. On the original website, the article has pictures and advertisements around it, which SAV does not consider to process. SAV will only process the plain text of the title and article.

SAV combines the use of multiple GIS tools and data sources to improve the user's visualization of the locations and relationships between locations on a map. There are two major components within SAV: The Parse, Identify, & Geo-Lookup (PIG) Java Class, and the Data Integration and Processing (DIP) toolset. PIG utilizes CLAVIN to parse the article text to identify locations and to perform geo-lookup. DIP integrates the geographic points that PIG identifies with the location boundaries and uses map layout templates to generate a map.



San Ramon Man Arrested After Allegedly Firing into Ex-Girlfriend's House

Craig O'Sullivan, 62, of San Ramon was arrested on Saturday night

By Autumn Johnson/ Editor (Patch Staff)
Updated March 17, 2014 at 12:05 pm

By Bay City News—

A man was arrested after allegedly firing shots into his ex-girlfriend's San Leandro home on Saturday night, police said.

Officers arrested 62-year-old San Ramon resident Craig O'Sullivan shortly after they responded at about 10:45 p.m. Saturday to a report of a shooting in the 1200 block of Pacific Avenue near Thrasher Park. Police arrived at the residence, located about two blocks from the San Leandro BART station, and determined that the victim's ex-boyfriend had come to her home and discharged a firearm.

O'Sullivan fled the scene prior to police arriving but officers caught up with him as he was driving on Marina Boulevard. He was arrested and booked into county jail in connection with the shooting, police said.

Figure 4 Example of a News Article to be Used in the Case study

The SAV process flow (Figure 5), highlights the major data inputs, processes, and the final data output of SAV. (Step1) The user copies the text from the news article, pastes it into a blank text document and saves the file. (Step 2) PIG calls NER functions from CLAVIN to parse the text file and identify locations. (Step 3) PIG also utilizes CLAVIN to perform geo-lookup of each location in a gazetteer to provide the geographic coordinates. PIG creates a Comma Separated Value (CSV) file and stores each location as a row in the file. (Step 4) The CSV file is an input to DIP to combine the points with the boundary of the corresponding locations using rules to choose the polygon that best matches the point. (Step 5) As an output, DIP creates the final map with the use of templates for format and style.

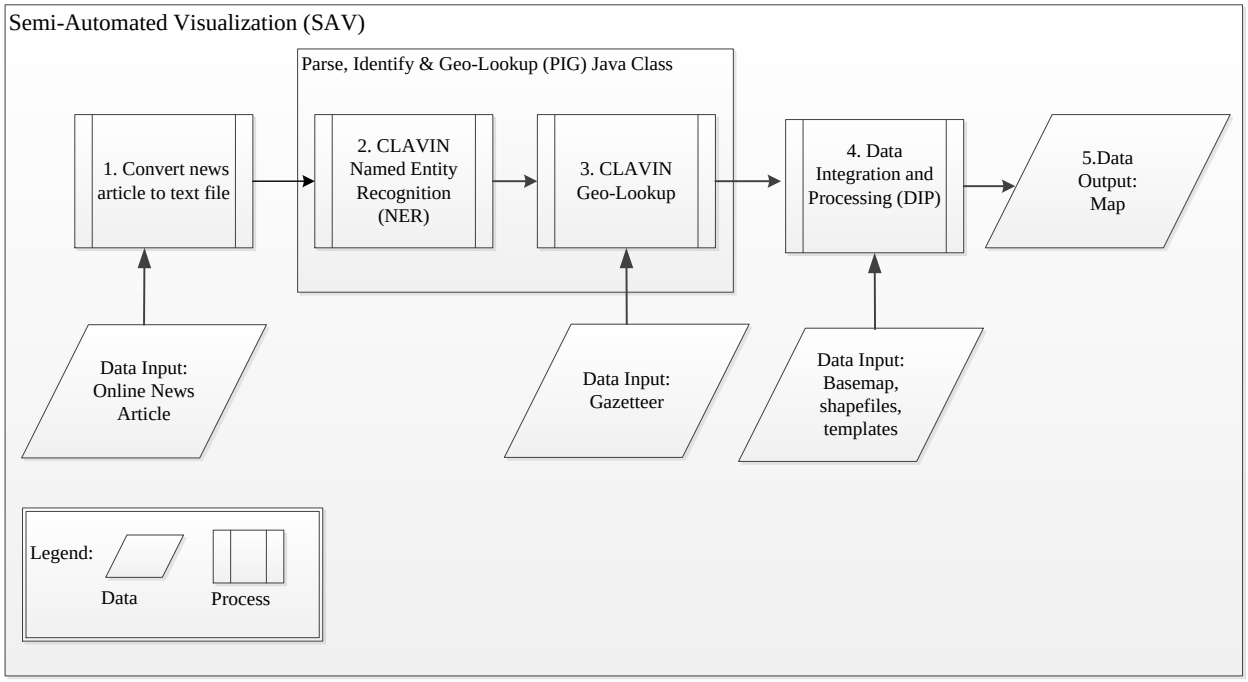


Figure 5 Semi-Automated Visualization (SAV) Process

The output of SAV is a map that uses the boundaries of a location as a representation of said location. The top image in Figure 6 displays the output of running the case study article though the CLAVIN online interface. Each location is a point. The image on the bottom of Figure 6 is a manually created map that shows the desired output of SAV. The benefits of the map on the bottom include illustration of the boundaries, labels on each location, and a view with more details than the map the CLAVIN online tool generates. The image also contains references to show the data source.

The remainder of this chapter is organized in sections to describe method, data, and processing components of SAV. Section 3.2 describes the study area that the case study uses to demonstrate the method of SAV. Section 3.3 describes the data,with sub-sections for article data, gazetteer data, and geographic data. Section 3.4, Semi-Automated Visualization Processing, is broken into two sections to describe the two main processing functions: PIG (Section 3.4.1) and the DIP toolset for data integration and map creation (Section 3.4.2).

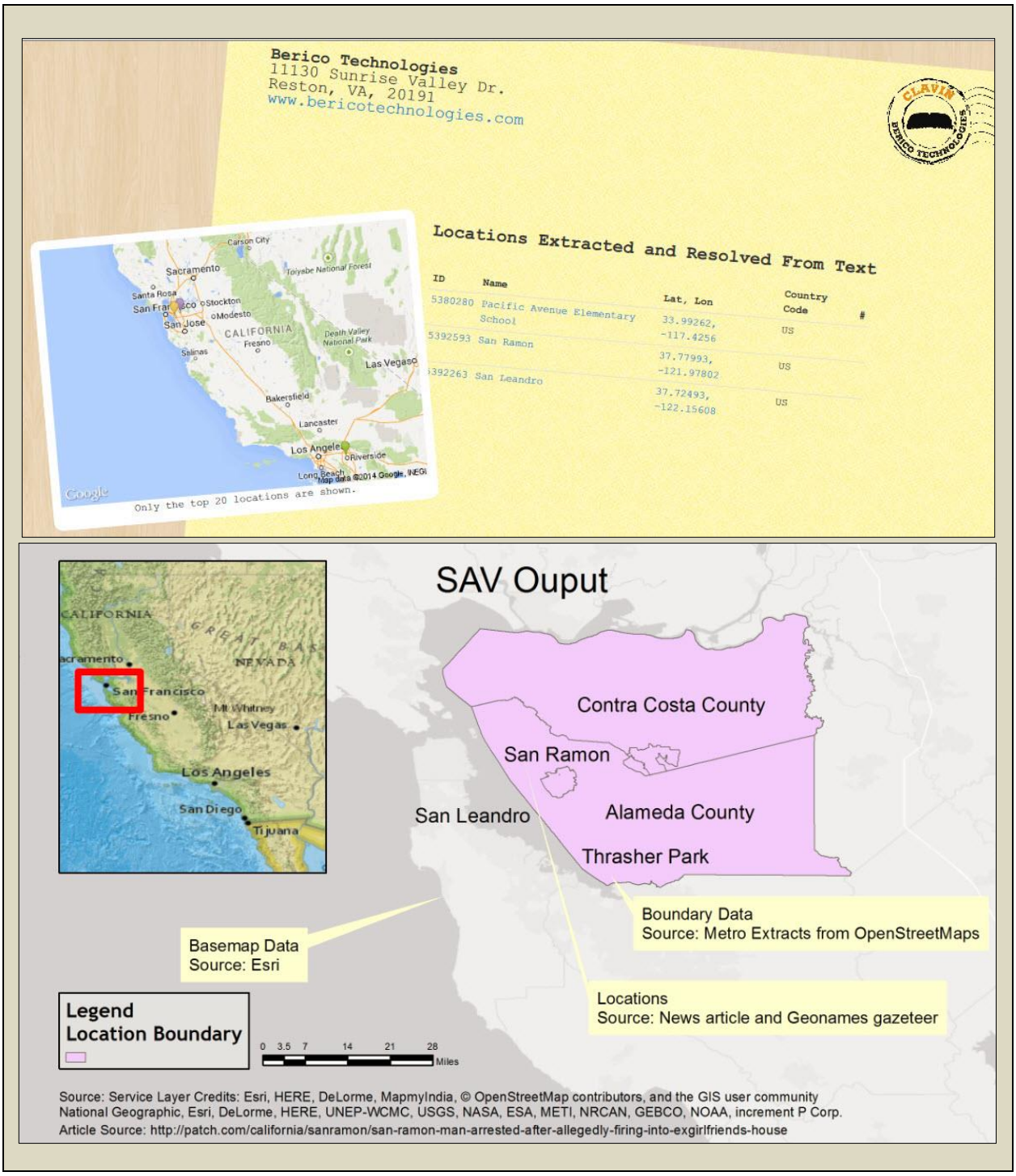


Figure 6 Top: CLAVIN Online Tool Map Bottom: Output of SAV Top Image by Berico Technologies (Berico Technologies 2014)

3.2. Study Area

The San Francisco Bay Area (Figure 7) in California, United States, was selected as the study area because it can be visualized at a reasonable scale that allows the relationships between locations to be visible. However, it is still a complex enough region to have a variety of location types: cities, counties, parks, and schools. The Bay Area, as it will further be referred to, is a highly populated region that has many different regional news publications to select news articles from.

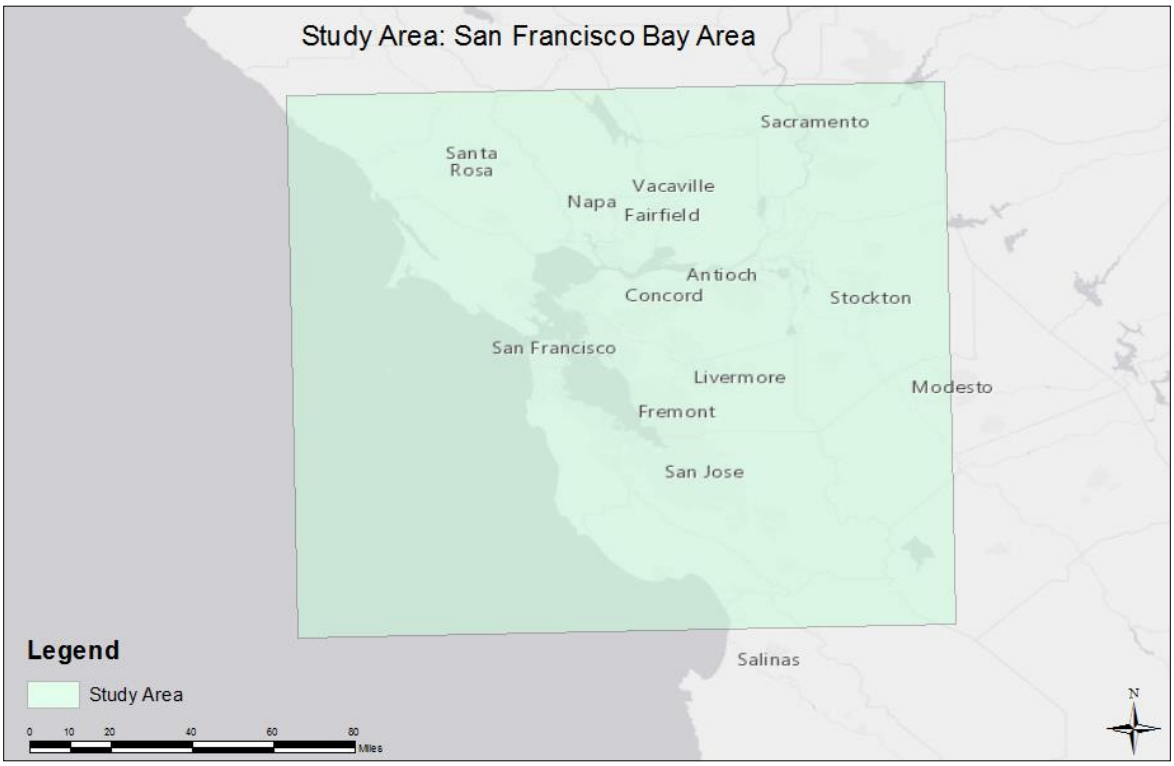


Figure 7 The Study Area: The San Francisco Bay Area

3.3. Data

SAV requires several data sources to support the application. The primary data sources that feed SAV are the news articles (3.3.1). SAV also requires a gazetteer database to provide geographical coordinates of each location PIG identifies (3.3.2). To map the locations, SAV uses

geographic data: a basemap and boundary representation of Bay Area locations (3.3.2). The combination of these data sources and some geo-processing will produce a complete map.

3.3.1. News Article data

Article data comes from online news sources. Each article needs to have multiple unique locations within the study area. Regional online news sources will likely mention locations within the region the study area more often than national and global news sources. *The Contra Costa Times*, *The Patch*, and *The San Jose Mercury News* are three regional news sources that are local to the study area and used as sources for news articles. Figure 8 is the case study article in a plain text format.

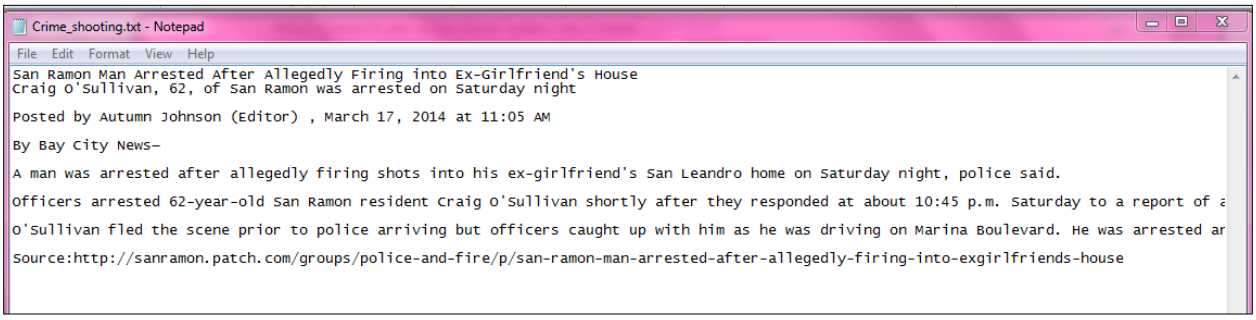


Figure 8 Article Text Manually Copied into a Text File

3.3.2. Geographic Data

PIG adds geographic context to the locations it extracts from the text file by looking up each of the geographic coordinates in a gazetteer. The gazetteer stores the location name, alternate location names, coordinates, and supporting attribute data for over 10 million locations. GeoNames.org (2014a) provides the AllCountries gazetteer that CLAVIN, the tool that is being used for geo-lookup, uses. The gazetteer uses the World Geodetic System 1984 (WGS84) coordinate system as the spatial reference for the geographic coordinates.

In order to define the boundaries of the locations, the application requires a reference polygon feature class that contains these boundaries. The source of the polygon feature class for the study area is Metro Extracts (Migurski 2014). Metro Extracts is a site that uses the OpenStreetMaps (OSM) as a source and provides extractions of metro areas throughout the globe. OSM is a community-driven open source data repository for GIS data (2014b). The polygon feature class' spatial reference is the WGS84 coordinate system, the same as the gazetteer data. The Bay Area polygon file contains the boundaries of about 200,000 locations including, but not limited to: cities, counties, parks, structures, mountains, and lakes. The file contains the shape and area of the polygon along with many attributes such as boundary, bridge, building, leisure, and military. The attribute only has a value if the feature is the type of feature mentioned in the attribute name.

ArcMap 10.2 provides built-in basemaps illustrating the foundational elements of the maps including country and state boundaries, and labels of those features. Basemaps are useful to provide the foundation of the map. The "Light Grey Canvas Map" from ArcGIS Online, provides a good basemap that does not distract the viewer from the location data on the map that is being displayed, it just provides the outline to natural features and administrative boundaries (Esri 2014b)

3.4. Semi-Automated Visualization Processing

SAV has two processing components that the following section describes. The case study article introduced earlier in this chapter is the input to SAV. The first component, PIG, is a custom Java Class that identifies locations in the article text and converts locations into geographic coordinates. The coordinate data feeds into the second component, DIP, which integrates the boundaries of locations to create a map using templates. The expected output at the

end of the process is a map that provides context to the locations and relationships that SAV identified. Figure 9 shows how the components of SAV integrate with one another. The data attributes, process flow, and case study data for these components demonstrates the functionality of these processes.

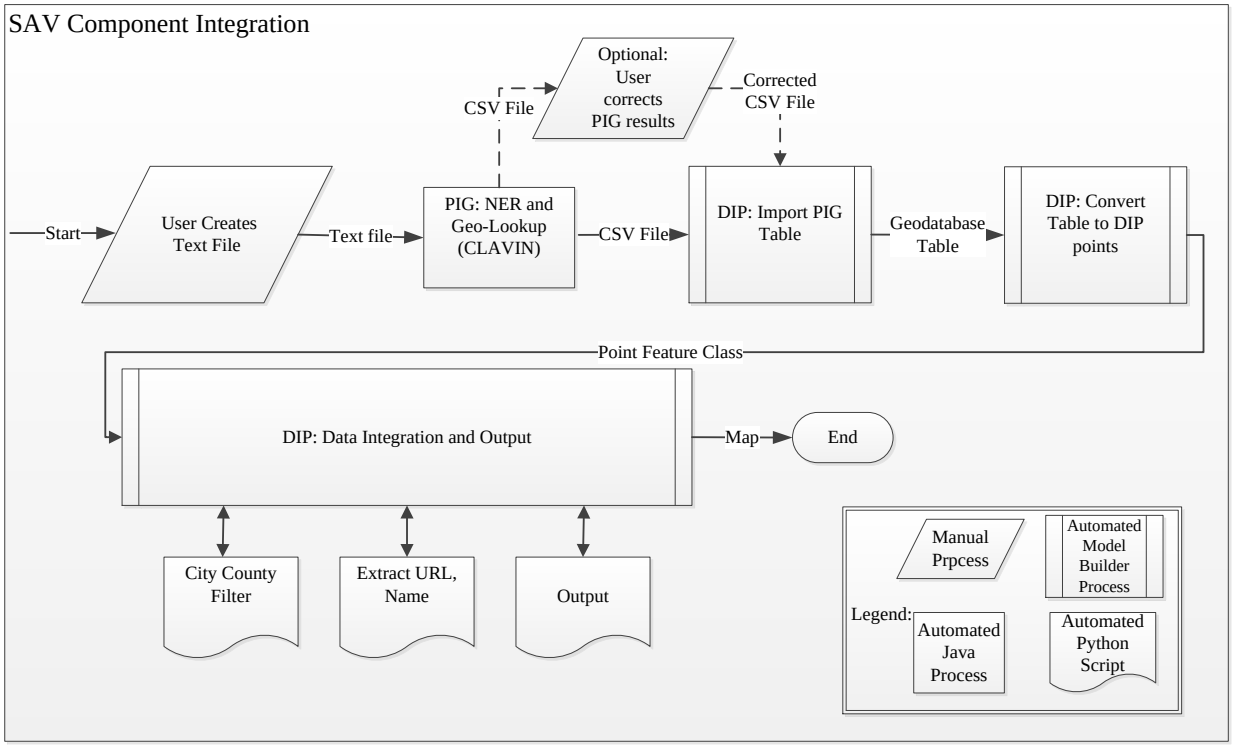


Figure 9 SAV Component Integration

3.4.1. Parse, Identify, and Geo-lookup (PIG)

NER is a very complex research problem on its own (Chapter 2). The approach was to use existing software for this portion of the work. PIG uses the parsing and geo-lookup functions of CLAVIN. PIG incorporates CLAVIN functionality into this thesis project with customization to allow processing multiple files in a single run and to save output in a file.

Prior to running PIG, the user must: select articles, create plain text copies, and store those copies in the input folder. Once the folder contains the plan text articles, the user can

execute the program. Figure 10 is a high-level process flow of the logic of PIG. (Step 1) PIG selects the first text file for processing. (Step2) PIG calls a CLAVIN text utility to convert the file to a string so that the program can parse the text. (Step 3) PIG parses the text and stores words that are locations in an array structure. During this process, CLAVIN has logic to resolve ambiguous locations, one of the issues discussed in Chapter 2. (Step 4) PIG iterates through the array structure of locations and assigns longitude and latitude coordinates from the gazetteer. (Step 5) The application stores the locations internally in an attribute. (Step 6) PIG creates a new CSV file to store the output. (Step 7) PIG loops through the array of locations and looks up the corresponding attributes in Table 1, providing the results as output into the CSV file, a list of rows. Each row has a location and details associated with a location, such as the longitude, latitude, feature class, feature code, country, admin1 code, and admin code 2. Once all locations are written to the file, it is closed (Step 8) and the file output is complete (Step 9). If there are more files in the input folder, the process will repeat until there are no files remaining.

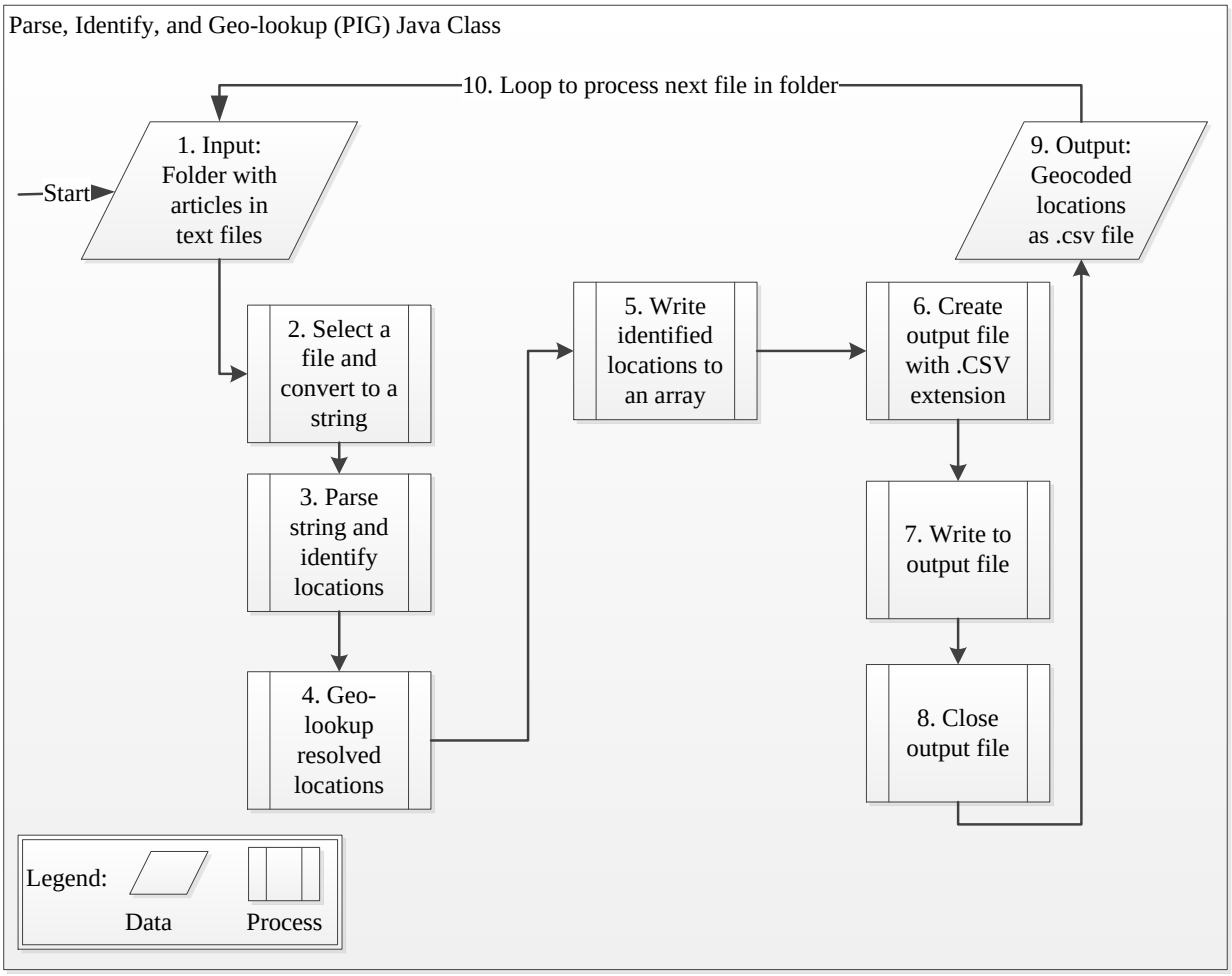


Figure 10 PIG Java Class Process Flow

Table 1 Location Attributes in PIG Output File

Attribute Name	Description
GeoNameID	A unique identifier of a record in the GeoNames database
ResolvedName	The location name that CLAVIN determines is the best match for the ArticleText
ArticleText	The text of a location reference as it originally appears in the news article
FeatClass	A classification of the type of location based on the classification of GeoNames.org
FeatCode	A sub-classification of FeatClass that is more specific based on the classification of GeoNames.org
X	Latitude of the resolved location in decimal degrees using WGS 84 coordinate system
Y	Longitude of the resolved location in decimal degrees using WGS 84 coordinate system
admin1Code	First administrative boundary corresponds to fipscode
admin2Code	A code for the second administrative division which corresponds to a county in the US
CountryCode	A two-letter country code based on the ISO-3166 2-letter standard

Source: Data adapted from GeoNames.org (2014d)

Figure 11 shows an example of the CSV file output of PIG using the case study article (Figure 8). The benefit of having the output in a CSV file is allowing DIP to process the information into a geodatabase table without any manual intervention.

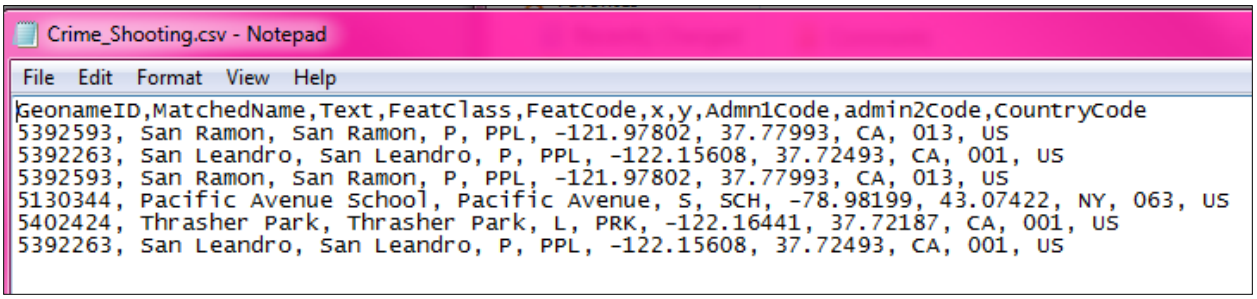


Figure 11 PIG CSV Output File For Case Study Article

3.4.2. Data Integration and Processing (DIP)

The DIP toolset relies on three tools built in ArcGIS Model Builder and three custom Python scripts to automate data integration and map output. The first two Model Builder tools perform pre-processing steps (Section 3.4.2.1). The purpose of the pre-processing tools is to create a point feature class in a geodatabase that has a single point for each location PIG extracts from the news articles. The Data Integration and Output (DIO) tool (Section 3.4.2.2) pulls all the previous work together to create a map.

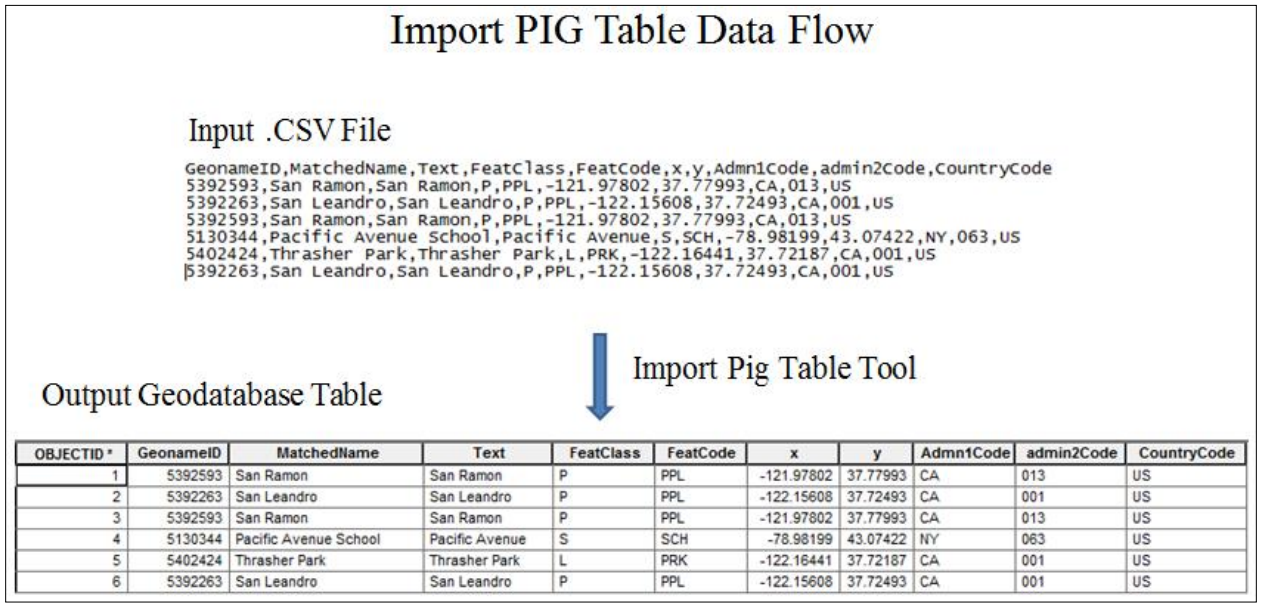
3.4.2.1. Pre-Processing

The first pre-processing tool is the Import PIG Table tool. This tool imports the CSV file that PIG generates into a geodatabase table. When a user executes the Import PIG Table tool, he receives a prompt to enter two input variables: Folder, the folder location that stores the CSV files and Geodatabase, and the name of the geodatabase the tool creates the new tables. The tool iterates through the input folder and creates a table for each CSV file and Table 2 summarizes its input and output parameters.

Table 2 Import PIG Table Data Parameters

Parameter	Parameter Type	Explanation	Data Type
Geodatabase	Input	The destination geodatabase where the tool creates the tables	Workspace
Folder	Input	Source folder for files PIG generates containing X,Y coordinates in a coma separated vales(CSV) format	Workspace
Table	Output	Table in destination Geodatabase	Table

The data flow for the Import PIG table tool is shown in Figure 12 below, using the case study data as an example. The tool imports the CSV file with six locations to create a geodatabase table with the same data. Each attribute from the CSV file has its own column in the geodatabase table.



OBJECTID *	GeonameID	MatchedName	Text	FeatClass	FeatCode	x	y	Admn1Code	admin2Code	CountryCode
1	5392593	San Ramon	San Ramon	P	PPL	-121.97802	37.77993	CA	013	US
2	5392263	San Leandro	San Leandro	P	PPL	-122.15608	37.72493	CA	001	US
3	5392593	San Ramon	San Ramon	P	PPL	-121.97802	37.77993	CA	013	US
4	5130344	Pacific Avenue School	Pacific Avenue	S	SCH	-78.98199	43.07422	NY	063	US
5	5402424	Thrasher Park	Thrasher Park	L	PRK	-122.16441	37.72187	CA	001	US
6	5392263	San Leandro	San Leandro	P	PPL	-122.15608	37.72493	CA	001	US

Figure 12 Case Study Data Input and Output Data Example

The next pre-processing tool is the Convert Table to DIP Points tool. This tool integrates the use of the Create Feature Class, Frequency, Delete Identical and Join tools that ArcGIS Model Builder provides. The purpose of this tool is to create a point feature class from the X,Y coordinates within the geodatabase table. The tool also prepares that data for map display by adding a new attribute frequency and eliminating the duplicate points. A point feature class with no duplicates and the frequency attribute is the final output of the Convert Table to DIP Points tool. Table 3 summarizes the parameters of the tool.

Table 3 Convert Table to DIP Points Data Parameters

Parameter	Parameter Type	Explanation	Data Type
Geodatabase	Input	A geodatabase that stores input tables and output point feature classes	Workspace
Spatial Reference	Input	The spatial reference of the coordinates in the input tables for (Example: WGS84)	Workspace
Point Feature Class	Output	Point feature class the tool creates that has no duplicates and a new attribute frequency	Feature Class

Figure 13 is a representation of the data flow using the case study data to describe the tool function. The user executes the tool and receives a prompt to enter the Geodatabase and the Spatial Reference parameters described in the table above. The tool converts the six locations in the input table into a point feature class using the spatial references the user provides. Within the Convert Table to DIP Points tool, the Frequency tool calculates the frequency of each location and stores it in a temporary table. Later on, the frequency is used to show the reader how many times the location was mentioned in order to highlight the importance of that place. Two locations in the case study data have multiple occurrences: San Ramon and San Leandro. The Delete Identical tool deletes the duplicates from the point feature class; in the case study data, this step removes duplicate entries for San Ramon and San Leandro. The Join Field tool unites the temporary frequency table with the point feature class.

Figure 14 below displays the output of the pre-processing tools using the case study data on a U.S. base-map. The next section will discuss data integration of the boundary data and how SAV creates the final map output with templates.

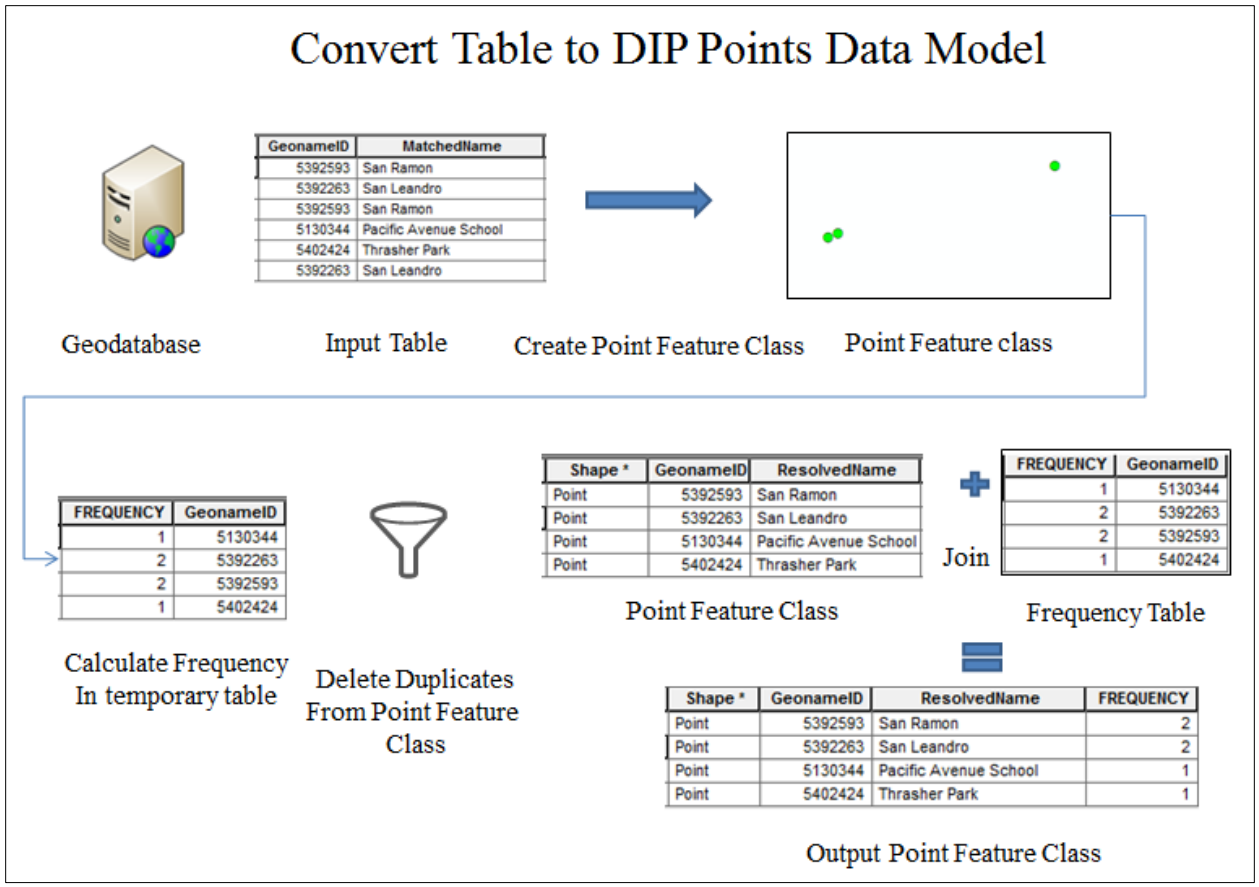


Figure 13 Example of the Convert Table to DIP Points Tool Case Study Data Flow

Figure 14 below displays the output of the pre-processing tools using the case study data on a US basemap. The next section will discuss data integration of the boundary data and how SAV creates the final map output with templates.

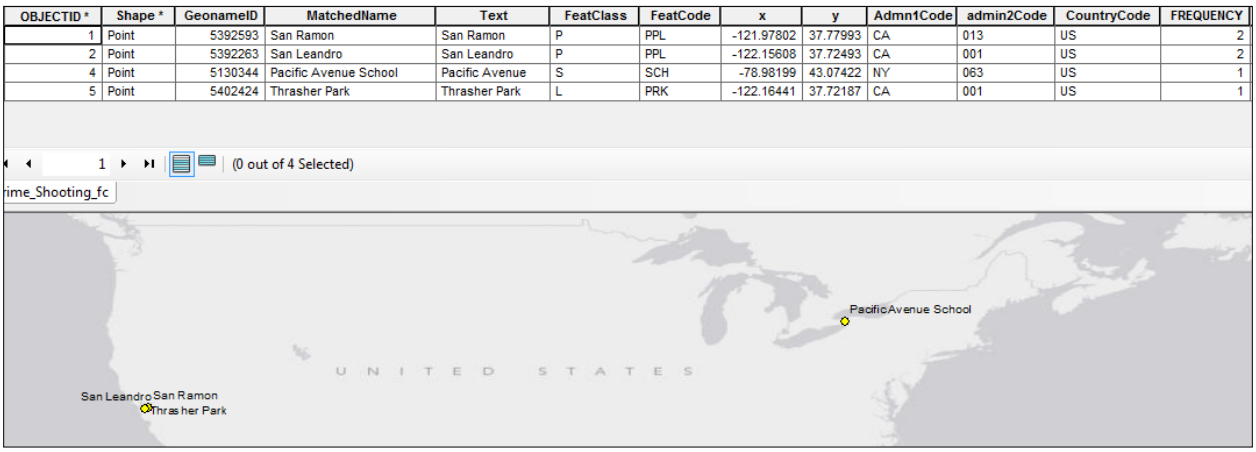


Figure 14 Case Study Output of Convert Table to DIP Points Tool

3.4.2.2. Data Integration and Output (DIO)

This section describes the DIO tool which contains eleven tools that ArcMap provides and three custom python scripts. There are seven data input parameters within the tool. Two input parameters are for the data integration portions, and the other five input parameters are strictly for the map creation. There are two outputs of the DIO tool. Table 4 summarizes the functions of the data parameters the DIO tool requires. These data parameters enable the tool to integrate the point feature class with a polygon data source and some layout and style templates in order to create a map.

Figure 15 is an example using the case study data on how the data flows though the tool to arrive at the final map results. The tool selects the point feature class from the geodatabase and performs a spatial join with the reference polygon feature class of the San Francisco Bay Area. The output of this process is a polygon feature class that contains seven polygon features. Not all of the results are relevant for display, so the tool needs logic to filter in order to find the polygons that are most likely to be the targeted results. The next section describes the filter logic in detail.

The output of the filter step is a polygon file that contains Thrasher Park and the two cities, San Leandro and San Ramon, that were in the original article. The polygon file also contains the counties that the cities are in, which are Alameda and Contra Costa respectively, to show the relationship between city and county. The polygons for Thrasher Park for the city and county will be removed systematically so the park is the only relevant polygon for that point. In addition, the tool filters the point class so only points that are not joined to a polygon are referenced as a point. This operation leaves Pacific Avenue School in the point file because it is outside the study area where there is no polygon data. A custom script prepares the article name

and URL of the article for display on the map. With all of the parameters complete, the custom Python script, OutputScript, runs to generate the final map (Section 3.4.2.4).

Table 4 Data Integration and Output Tool Data Parameters

Parameter	Parameter Type	Explanation	Data Type
Geodatabase	Input	Contains the point feature classes which are input parameters and stores the output feature classes	Workspace
Polygon	Input	A reference polygon file that contains all the boundaries of locations within the study area	Feature Layer
Folder	Input	Where the final maps will be output and stored	Folder
Point_Template	Input	A point layer file that has the symbology and label styles that the toolset uses as a template for new point feature classes	Layer File
Polygon_Template	Input	A polygon layer file that has the symbology and label styles that the toolset uses as a template for new polygon feature classes	Layer File
MapLayout	Input	ArcMap Map document that is a template for the format of the final map output including data frame placement	MXD
Input	Input	Folder where the URL's.dbf tables is stored;table contains the article name and URL names and is used to display the article and URL name on the map output	Folder
Map Image	Output	Final output of a .PDF map that cannot be edited	PDF
ArcMap Document	Output	Final image output of which can be used to edit the final map output if desired	MXD

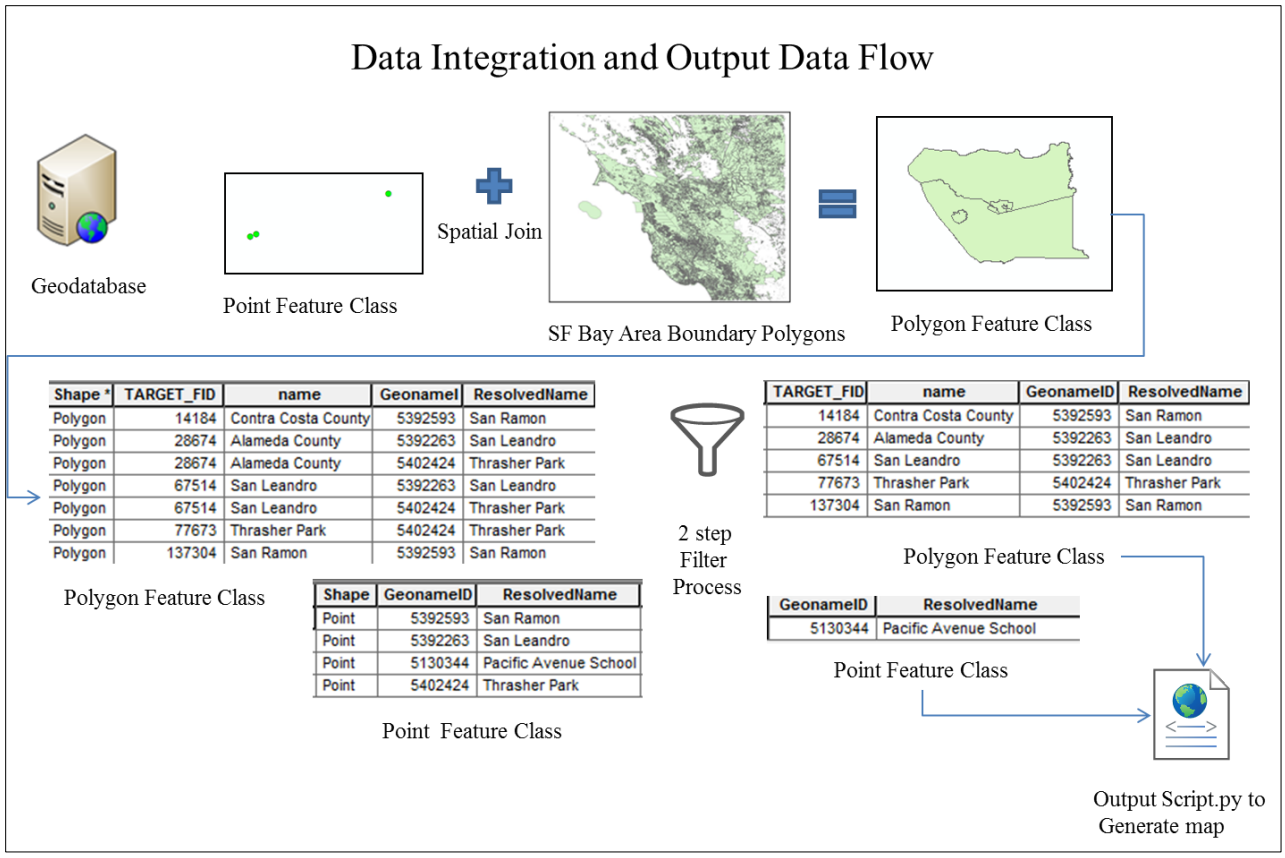


Figure 15 Case Study Data Flow Through the DIO Tool

3.4.2.3. Filter Criteria

The filtering criteria the DIO tool uses is a unique feature of SAV. The Spatial Join tool that DIO uses to integrate point and polygon data will join several polygons to a single point. For display purposes, filters will remove the polygons that are not the best matches. The rules for filtering are specific to the attributes that are in the point and polygon feature classes. If the data sources or display requirements change, the rules would need to be re-examined. For this thesis project, each point in the study area should correspond to one polygon unless it is a city. A city point will correspond to a city and county polygon. This is an example of how SAV can display spatial relationships. Further research and experimentation may change the rules to show different relationships.

The Geonames data set, what the point feature class is based on, contains the Feature Class and Feature Code attributes. These contain codes that correspond to detailed classifications of the location types. For example, Feature class ‘A’ refers to an administrative boundary. There are over 20 sub-classifications in the Feature Code attribute that correspond to various specific varations of an administrative boundary. The classification structure is complex and contains many subdivisions. For simplicity throughout this thesis project; the Feature Class ‘A’ will correspond to a county. Feature Class ‘P’ will refer to a city. Counties and cities are commonly identified features within CLAVIN. This is a generalization that may not apply in all cases, but is simplified for the sake of explanation.

The polygon attribute data is not as specific as the point data; while there are many attributes, the majority of polygons have no values for any of these. The one that remains consistent is the boundary attribute. If the boundary of a polygon has the value of ‘administrative’, then it corresponds to an administrative boundary such as a city or county. If the boundary attribute is not ‘administrative,’ it is another feature, man-made or natural, such as a church, school, lake, or mountain. The DIO tool has rules that use the boundary attribute of the polygon feature and the Feature Class attribute of the point class. Table 5 below summarizes these rules and point and polygon attributes the rules use. There are many classifications in the Feature Class attribute but the rules are limited to include structures, parks, administrative boundaries and natural features. These are features that CLAVIN recognizes and are prevalent in news stories.

Table 5 Rules to Filter Polygons to get the Most Likely Matches

Rule	Point Feature Class value	Polygon Boundary attribute value
A point that is a county should be joined to a polygon that is an administrative boundary	A for County	Administrative
A point that is a city/town should be joined to a polygon that is an administrative boundary	P for City/town	Administrative
A point that is a park should not be joined to an administrative boundary	L for park	Not Administrative
A point that is a natural feature (i.e. beach, river) should not be joined to an administrative boundary	T for natural feature	Not Administrative
A point that is a structure (i.e. school, church, airport)should not be joined to an administrative boundary	S for structure	Not Administrative

The Select tool within the DIO tool has an SQL statement using the rules above to filter the results of the Spatial Join tool. The output of the Select tool is a polygon feature class that contains the polygons spatially joined to the points that meet the attribute rules. If a point is a structure, park, or natural feature and it is joined to the city or county polygon that it resides in, the city and county polygons will not be selected because the point type is non-administrative. This way the tool eliminates unnecessary records and the polygons that remain are likely to be the correct result. Using the rules, the Select filter will remove the rows in Figure 16 shown with blue highlights. The polygon for Alameda County and San Leandro related to Thrasher Park is not a relevant location for display.

Polygon Location	Polygon Bounda	Point Location	Point Feat Class
Contra Costa County	administrative	San Ramon	P
Alameda County	administrative	San Leandro	P
Alameda County	administrative	Thrasher Park	L
San Leandro	administrative	San Leandro	P
San Leandro	administrative	Thrasher Park	L
Thrasher Park		Thrasher Park	L
San Ramon	administrative	San Ramon	P

Figure 16 Select Filter Case Study Data

Administrative boundaries are not easy to filter because the polygons do not distinguish between the different types of boundaries. A custom CityCountyFilter python script provides the second stage of filtering for administrative boundaries. The logic is rule-based and summarized in Table 6. The script assumes that if a point is joined to more than one polygon of an administrative type, there is probably a city and county referenced. In the case study, there are only two levels to this hierarchy structure because within the size of the study area there are no state or country polygons. The tool uses the area of the multiple polygons joined to a single point to decide which polygon is the best match. This rule assumes that if a city is within a county, the city will have a smaller area than the county. This assumption may be over-simplified since the Feature Classes cover more types of locations than just cities and counties. Since the polygon's attributes are limited, it seems to be the only available option to filter the data further. Chapter 4 will discuss the results using this logic.

The CityCountyFilter script compares each feature against all other features using nested loops and runs conditional checks to see if a feature should be filtered out. In the CityCountyFilter Python script if the point Feature Class is 'A,' that means the original point is a county and the polygon with the larger area is likely the correct result, the smaller polygon is

likely incorrect and should be deleted. The tool removes the smaller polygon in this case because it is likely a random city within the county and is not relevant to the news article.

Table 6 Summary of Rules for the CityCountyFilter Tool

Rule	Point attribute	Polygon Attribute	Logic	Action
The original point is a county select the larger polygon	Feature Class = A for county	way_area	if feature 2 way_area < feature 1 way_area	delete feature 2
The original point is a city select the smaller polygon	Feature Class = P for City	way_area	if feature 2 way_area > Feature 1 way_area	set feature 2 boundary = county set feature 2 Frequency = 0
All other conditions				do nothing

Conversely, if the original point was a city, then the polygon with the smaller area is likely the better match. In the case of a city, it may be relevant to also know what county the city is in. The relationship between city and county provides more spatial context in this illustration. Therefore if the point associated with two polygon features is equal and the Feature Class is ‘P,’ then the polygon with the smaller area is the best result and no attributes change.

In this scenario, the polygon with the larger area can be kept, but the tool updates some of the attributes in order to distinguish a county that is matched with a city point from a county matched to a county point. The frequency attribute is set to zero because the polygon was not actually in the text, it just provides contextual information. The boundary attribute is also set to county so that there is some distinction between city and county features. If a pair of features does not meet any of these criteria, nothing changes and the script iterates to the next combination until the script compares all pairs. Figure 17 displays the output of the

CityCountyFilter script. The polygon boundary and frequency attributes for Alameda and Contra Costa counties are now set to county and 0 so these counties can be displayed differently than a county that was actually mentioned in the text.

Polygon Location	Polygon Boundary	Polygon Area	Point Locations	FeatClass	FREQUENCY
Alameda County	county	0.217025	San Leandro	P	0
San Leandro	administrative	0.004094	San Leandro	P	2
Contra Costa County	county	0.213161	San Ramon	P	0
San Ramon	administrative	0.00493	San Ramon	P	2
Thrasher Park		0.000001	Thrasher Park	L	1

Figure 17 Case Study City County Filter Results

It is important to note that even if the tool deletes a county feature, in this step it may still appear in the final result because the county itself may have also been in the text. In this case, the county polygon will be joined to the point that corresponds to the county. The polygon feature class is complete. If no polygon match was found for a point during the spatial join and filtering operations, then the locations should be displayed as the original point so that no locations from the text are missing. The tool deletes the point in the point features class that has a polygon match. The result is a features class with the relevant polygons and a point feature class with the points that did not have a polygon match.

3.4.2.4. Output

The OutputScript is where all the previous work comes together and the tool generates a map. Figure 18 helps summarize what the script does. The MapLayout template, Point_Template, Polygon_Template and the one-line table with article name and URL, are all input along with the final point and polygon feature classes. The script opens the MapLayout template and replaces the data source of the Polygon_Template layer in the template with the polygon feature class and assigns that layer to the main data frame. A similar process is followed

for the point feature class using the Point_Template. The one exception is the assigning of the point layer to the alternate data frame because the features that are points are not the focus. The extent for this data frame needs to be set based on the data in the layer. Finally, the map document is saved as a new MXD file and exported to a PDF to the folder indicated by the folder variable. The variables used by the script are deleted and the DIO tool goes back to the iterator to see if there are any other point feature classes to process.

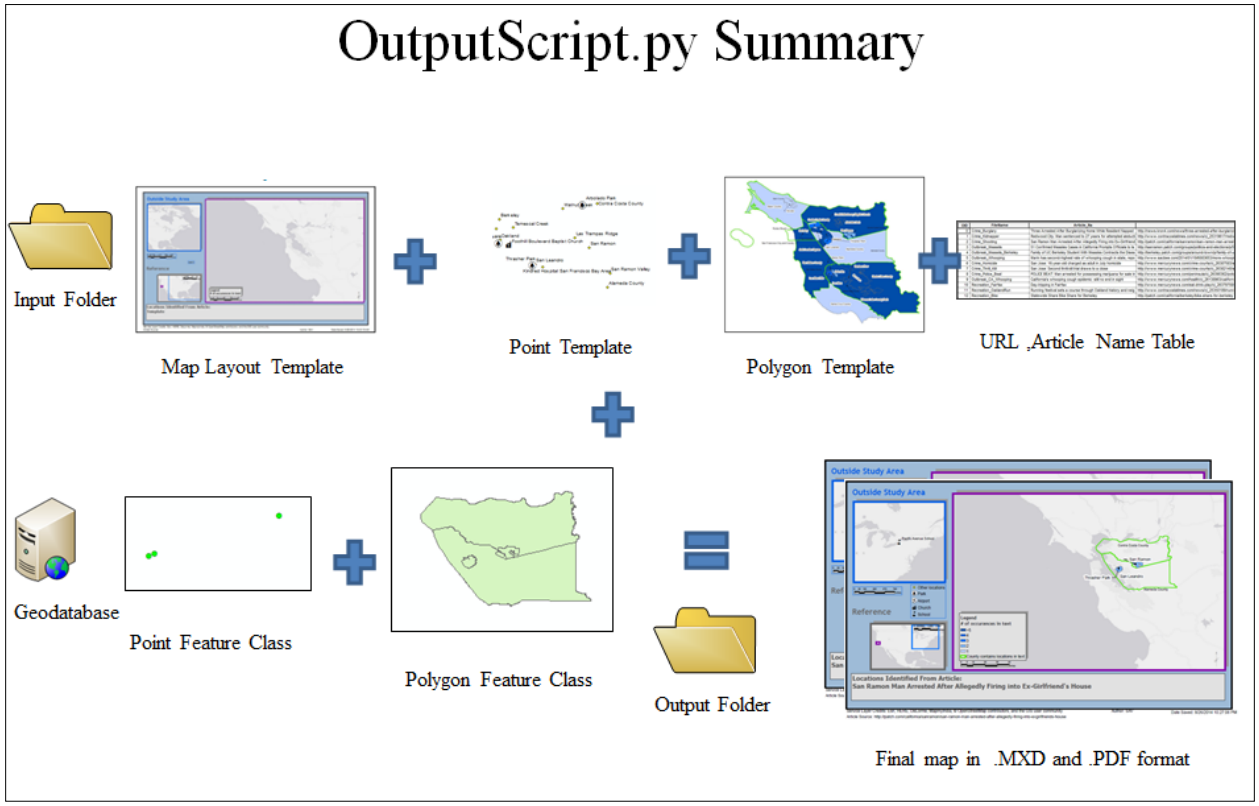


Figure 18 OutputScript.py Summary Data Flow

The tool generates the map based on the format in the map layout template, the symbology and label styles in the point, and the polygon template layers. The final map output is shown below in Figure 19.

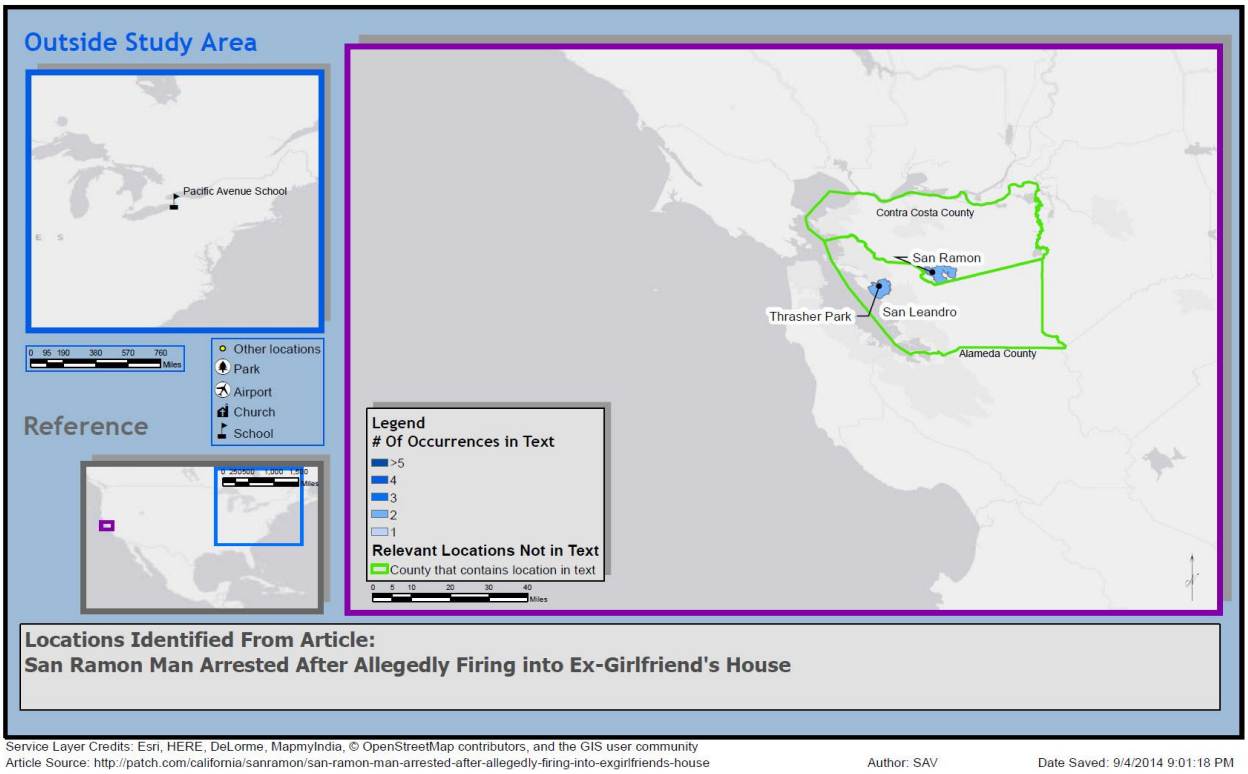


Figure 19 Final Map Output for Case Study Article

This chapter described the data sources and the method that is used for this thesis project using a case study of a news article to demonstrate the method. The next chapter discusses the experiment using this toolset and evaluates the results of the experiment and the effectiveness of the process.

CHAPTER FOUR: EXPERIMENT AND EVALUATION

This chapter describes an experiment that tests the effectiveness of SAV. It includes a description of the evaluation techniques, results of the experiment, and evaluation of the results. Section 4.1 describes the details of the article selection for the experiment and its design. Section 4.2 describes the techniques to evaluate the experiment results. Section 4.3 articulates the results of the experiment. The chapter closes with Section 4.4, an explanation of the survey design, results, and discussion.

4.1. Experiment Article Set

To test the usefulness of SAV, a set of articles was selected and run through the entire process. The data files and evaluation results described in this chapter are available on GitHub (https://github.com/spatial-computing/sav). The selection criteria used to find articles for the experiment were identifying articles with at least two locations in the study area and articles about crime, disease outbreak, or recreation.

Articles were selected in these three separate categories to determine if SAV is more useful for one type of news story over another. The first category was articles about crime because the articles often reference several locations. For an article about a crime, understanding the context and proximity of multiple locations could also be important to readers to understand how the events occurred. Articles about public health, or disease outbreaks, generally describe the location where the disease is a threat and may include locations where a contaminated person has traveled. Finally, a third category that benefits from spatial relationships, is recreation. There are a number of articles about hiking trails, running, tourist attractions, and other leisure activities that often mention a string of locations in one general area.

Three articles that met the minimum location requirements within the study area were selected for each category for the experiment to determine how well SAV could generate a map. Table 7 below shows the nine articles selected for the experiment, references to the article text in Appendix B, and short article names that will be used throughout this chapter to reference the news pieces. Appendix B contains the text of the articles that Table 7 in the same order.

Table 7 Table of Articles Selected for Experiment

Refernce to article text in Appendix B	Category	Short Article Name	Article Name
Figure B- 1	Crime	Hate Crime	*Hate Crime Charge in Palo Alto Beating Case*
Figure B- 2	Crime	Homicide	*San Jose: 16-year-old charged as adult in July homicide*
Figure B- 3	Crime	Shooting	*San Ramon Man Arrested After Allegedly Firing into Ex-Girlfriend's House*
Figure B- 4	Outbreak	Measles	*51 Confirmed Measles Cases in California Prompts Officials to Issue Warning*
Figure B- 5	Outbreak	Measles Berkeley	*Family of UC Berkeley Student With Measles Contracts the Disease*
Figure B- 6	Outbreak	Whooping	*Marin has second-highest rate of whooping cough in state, report says*
Figure B- 7	Recreation	Food	*Oakland chef wins Esquire's TV's 'Knife Fight'; Mixt comes to the South Bay; Jack's Oyster Bar opens in*
Figure B- 8	Recreation	Helix	*Helix: A mini Exploratorium for the South Bay*
Figure B- 9	Recreation	IH Trail	*Pleasanton: New Iron Horse Trail segment opens*

4.2. Evaluation Methods

The evaluation of this thesis project includes three parts in order to determine the following criteria: the quality of the data, the quality of the map output, and the effectiveness of the output. An evaluation of the quality of the NER results from PIG helps to gain an understanding of the data quality of the output. If the results from the PIG tool are poor, then it is likely the final map is inaccurate. Evaluating the map output is important to discover if SAV was

effective at mapping locations. A survey provides feedback from others to help determine how relevant the final maps are.

The author evaluated the PIG results by manually reading each article, highlighting the locations in the article, and comparing the highlighted locations with the PIG identified locations. The precision and recall of the PIG results were calculated to provide a quantitative metric to understand the quality of the results. Precision is the percentage of locations identified that are relevant, and recall is the percentage of relevant locations that are identified (Nikolajevs and Jekabsons 2013). In this case, relevant locations are the types of locations that exist in the GeoNames database, including: countries, states, counties, cities\towns, schools, and parks. Other features such as roads and rivers are not included as a relevant location in this study.

The list of locations manually identified, known as the ground truth, were compared against the PIG identified locations. Each location was classified as a true positive or a false positive. A true positive is a location that matches the manually identified locations, a false positive is a location that was incorrectly identified. For example, the article mentions the city of Richmond; the author knows based on the context that the article is referring to the Bay Area city in California. The context of the article makes it clear that Richmond is not the large city in Virginia. If PIG selected the city in Virginia, then this location would be a false positive.

Another variation of false positives is the case where PIG would identify a location that the author did not deem to be a location. For instance, Pacific Avenue was in one of the articles, but PIG does not recognize street names. Because there was an entry in the gazetteer for Pacific Avenue School, PIG matched the school with this location incorrectly. In the evaluation, this is an example of a false positive. Once the classifications of each location were made, the precision and recall metrics were calculated. Precision is the true positive divided by the sum of true

positive and false positives. Recall is the true positive divided by the ground truth. These metrics provide results as a percentage which illustrates a view as to how well PIG performed. The higher the percentage, the better the results are.

4.3. Quantitative Experiment Results

This section describes the results and evaluations of the experiment results. The results from PIG's use of CLAVIN may not be completely accurate because the NER and geo-lookup functions are areas that still require research. Due to the fact the precision and recall for the identification of the correct locations was relatively low (Table 8), the approach was to manually correct the results from PIG before DIP creates the maps. Manually correcting the PIG results allows the viewer to evaluate the map output without the distraction of incorrect locations.

DIP was executed with the articles and with the results from PIG as-is. The second execution of DIP was run with the results from PIG manually altered to resolve the ambiguous locations correctly and add any missing locations. This section will show the results from both runs of SAV and highlight the benefits and challenges of each run.

Table 8 Precision and Recall Metrics for Experiment Articles

Article Short Name	Precision	Recall	Ground Truth (Manually Identified Locations)
Hate Crime	100%	71%	7
Homicide	88%	78%	9
Shooting	83%	83%	6
Measles	100%	81%	16
Measles Berkeley	75%	25%	12
Whooping Cough	100%	33%	18
Food	82%	69%	11
Helix	91%	91%	26
IH Trail	88%	50%	14

The first article was a crime article about a beating case that was later described as a hate crime. Figure A-1the first run with the unaltered PIG results and Figure A-2 is the map with the manually edited PIG results. The only difference between the maps is that the original PIG results did not include Redwood City. The map generated with the manually-altered PIG locations includes the polygon from Redwood City and the county it is in, San Mateo County. SAV correctly mapped all the locations it identified with the original PIG results; it was just missing the locations that PIG did not identify. For the map generated with manually corrected PIG results, SAV was able to accurately map all the locations identified in the text. The precision for this article was 100% and the recall was 71%.

The next article is about a homicide in San Jose. In this case, the PIG results include an incorrect point in New Mexico for East San Jose Elementary shown in Figure A-3. Since in the GeoNames online database there is no East San Jose point to describe the region in the city, it was manually reassigned to San Jose for the manually corrected PIG results. In Figure A-4, the only other manual correction is that the city of San Jose was mentioned one more time than PIG recognized. This article also had fairly high precision and recall, over 75% for both (Table 8), so there were not very many manual corrections. SAV correctly mapped all the locations that PIG originally identified and the manually corrected locations.

The PIG results for the next article incorrectly recognized the street, Pacific Avenue, to be Pacific Avenue School, as shown in the Outside Study Area data frame in Figure A-5. Since PIG does not identify streets, there was no corresponding location for Pacific Avenue in the manually corrected PIG results shown in Figure A-6. There was also one occurrence of the city of San Ramon that was not recognized by PIG. The precision and recall for this article were both

83%. SAV correctly mapped all the locations according to the rules when the PIG locations were manually corrected.

The next set of articles covered disease outbreaks. The first example is based on an article about the relatively high number of measles cases in the Bay Area. Since the gazetteer does not have a complete listing of regions, such as Southern California, the reference was incorrectly associated with Southern California Logistics Airport (Figure A-7); the manually corrected location was just California. PIG also did not recognize Monterey County as a county; that was added to the manually corrected results as can be seen in Figure A-8 in the Outside Study Area data frame. The precision for this article was 100% and the recall was 81%, therefore there were not a lot of manual corrections. SAV generated all of the locations according to the rules and they matched the correct locations for both maps.

In the case of the article about the family of the UC Berkeley student that had measles, the PIG results were missing a lot of locations. The city of Berkeley was not recognized and there were five references to it. The city of El Cerrito was also not identified along with several references to Contra Costa County and a reference to California. Therefore, in this case, the map with the PIG results as-is in Figure A-9 is very different than the map with manually corrected results in Figure A-10. The recall metric was for this article was very low, at 25%, and the precision was 75%. These metrics clearly show that the results for PIG were poor in this case. SAV correctly mapped the locations PIG identified, whether they are accurate or not, and was able to correctly map 100% of the manually corrected locations.

The next example is interesting because the manually corrected results actually performed poorly in mapping. The original PIG results were missing seven occurrences of the city of Marin, two instances of Marin County and one instance of the city of Greenbrae so the

coordinates for these locations were added to the manually corrected results and run through the remaining SAV processes (Figure A-11). In this case, it was a data quality issue on the SF Bay Area Polygon file; there were no polygons for the city of Marin or Greenbrae. The city of Marin was just represented by the Marin County polygon since the city polygon did not exist. Greenbrae was correctly matched to the polygon for Marin County, and with the polygon for the city of Kentfield which is an incorrect match (Figure A-12). This may be because according to *Marin Magazine* (Jewett 2014), Greenbrae and Marin City are unincorporated areas of Marin County that the SF Bay Area polygon feature class does not appear to give names to or provide an attribute that would be useful for identifying these unincorporated areas.

The metric's low recall value of 33% is indicative of the fact that in this case several locations were identified incorrectly. In this case, the map with the PIG generated results may have produced a better result by leaving out that location. Mapping incorrect locations may cause confusion. This problem could be remedied by modifying the source polygon file to correctly describe the regions that are unincorporated portions of the county. Rules can be developed to handle these regions, or they could be omitted from the location extraction process (as PIG did) because they are unincorporated areas.

The next article (given a short title of Food) is about some local restaurants and mentions twenty-six locations. The PIG results were fairly accurate with a few wrong identifications for locations Santana Row, Santa Clara, and South Bay, all manually corrected. The locations originally identified are shown in the map in Figure A-13 and the manually updated locations are mapped in Figure A-14.

The main objective is to notice that with this article both maps of the city of San Francisco are incorrectly illustrated. The SF Bay Area Polygon Feature Class has one feature

called ‘San Francisco City and County’ and another feature called ‘San Francisco,’ which happens to have a larger area than the feature that has ‘County’ in the title. This appears to be a data quality issue with the SF Bay Area Polygon feature class. The ‘San Francisco’ polygon includes a large area in the bay which encompasses several islands that are part of San Francisco, the ‘San Francisco City and County’ polygon feature does not include these islands, so the area for that feature is smaller, which results in the incorrect locations to be selected to represent the city. This is an issue with the data and would not make sense to make a systematic adjustment to account for this. Other than this anomaly, SAV correctly mapped the PIG generated and manually corrected locations. The precision of the PIG results was 82% and the recall was 69%.

The article about the Helix Museum has two things to note. The location Carmel, Indiana was incorrectly selected by PIG (Figure A-15), the manually corrected city is Carmel-by-the-Sea in California (Figure A-16). This article also referenced the city of San Francisco, so there are the same issues that were mentioned in the previous article. Other than that, PIG and SAV performed as expected. The precision and recall were both 91%.

For the final article about new construction on the Iron Horse trail, the main difference between the maps is that the city of Dublin in this article is actually in California and not the well-known one in Ireland. The original locations are shown in Figure A-17 and the corrected locations are mapped in Figure A-18. In addition to a mislabeling of the city of Dublin, the manually corrected maps contain more references to the city of Pleasanton, Contra Costa County, and the city of Concord.

Overall SAV, including the rules developed to be used by SAV, performed well. Six of the nine manually corrected PIG results maps were completely accurate. Three of the manually corrected maps had four instances where locations were incorrectly resolved. Two of the errors

were attributed to a data inconsistency with the city of San Francisco, which causes the city polygon to have a larger area; it was displayed as the county, and the county polygon was displayed as the city. This could be rectified by correcting the source data. The other mapping issue occurred with two locations that are unincorporated parts of Marin County that were not correctly mapped because the SF Bay Area Polygon feature class does not contain unincorporated areas, or the attribute data was not clear enough to identify them. The solution here would be to correct the data in the source data file or refrain from considering places that are unincorporated areas within counties. Other than the two issues that were noted, SAV performed well and correctly mapped the locations according to the rules defined. The next section will cover the results of a survey designed to capture other people's opinions about the effectiveness and usefulness of the SAV developed maps.

4.4. Survey Design, Results, and Discussion

A survey was prepared using Survey Monkey, an online survey-building solution that provides tools to build surveys and collect and analyze results (SurveyMonkey 2014). The main purpose of the survey was to evaluate if the maps generated by SAV provided better visualization of locations and relationships in news articles than the existing text-to-map solution, the CLAVIN online tool. The respondents were asked to read the nine articles from the experiment and evaluate two maps for each article: the SAV map with manually corrected PIG results and the CLAVIN online tool map as seen below (Figure 20).

4.4.1. Survey Design

The respondents were asked three comparative questions to choose which of the two maps was more useful. In addition, for each article, the respondents were asked if having a map was useful, and were also provided a free form comments field for additional thoughts. The

survey was posed on the author's Facebook account and was made accessible to adults in her "friends list" which includes 270 people, ranging from close friends and family to school and work acquaintances. The survey was left open for one week for data collection. The respondents self-selected to opt-in to the survey. There were no incentives offered for survey completion.

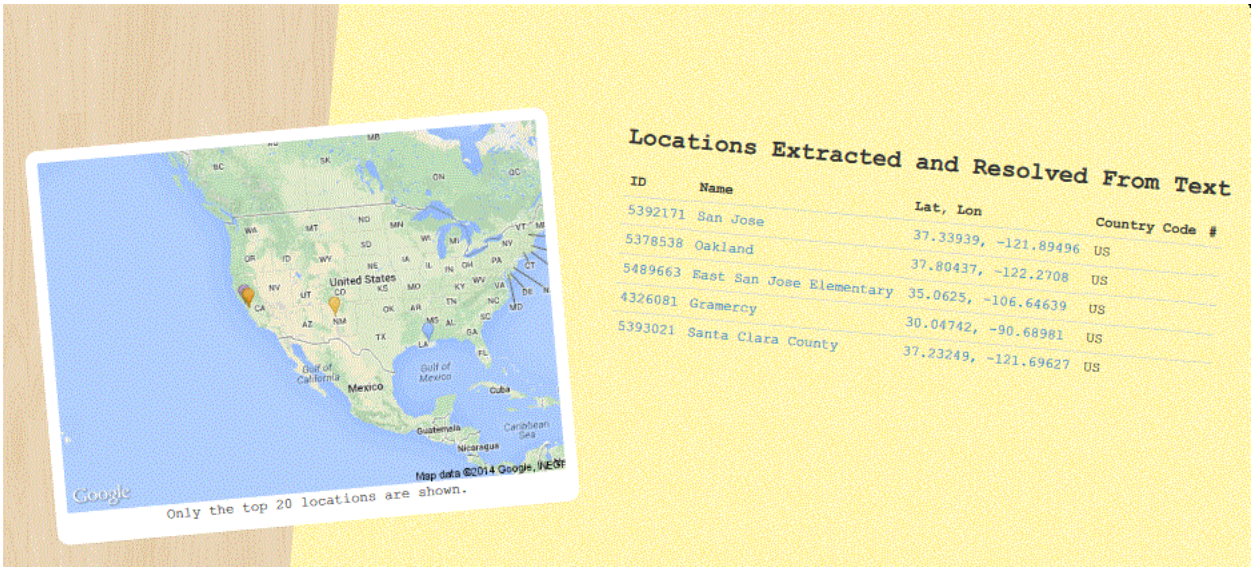


Figure 20 Example of a Map Generated by the CLAVIN Online Tool
Image by Berico Technologies (Berico Technologies 2014)

The first set of questions were designed to gain a little understanding about the person that was responding to the survey. Figure 21 below shows the questions and their options. The first question asks for the respondent's level of education, in order to have some background for interpreting the results. The second question was to find out the respondent's zip code; this was important in order to compare the responses of people that live within the study area with those that live outside the study area. Finally, the last question about the users was to find out how much GIS experience they have; this question was designed to see if the person's level of GIS experience had any correlation to his responses.

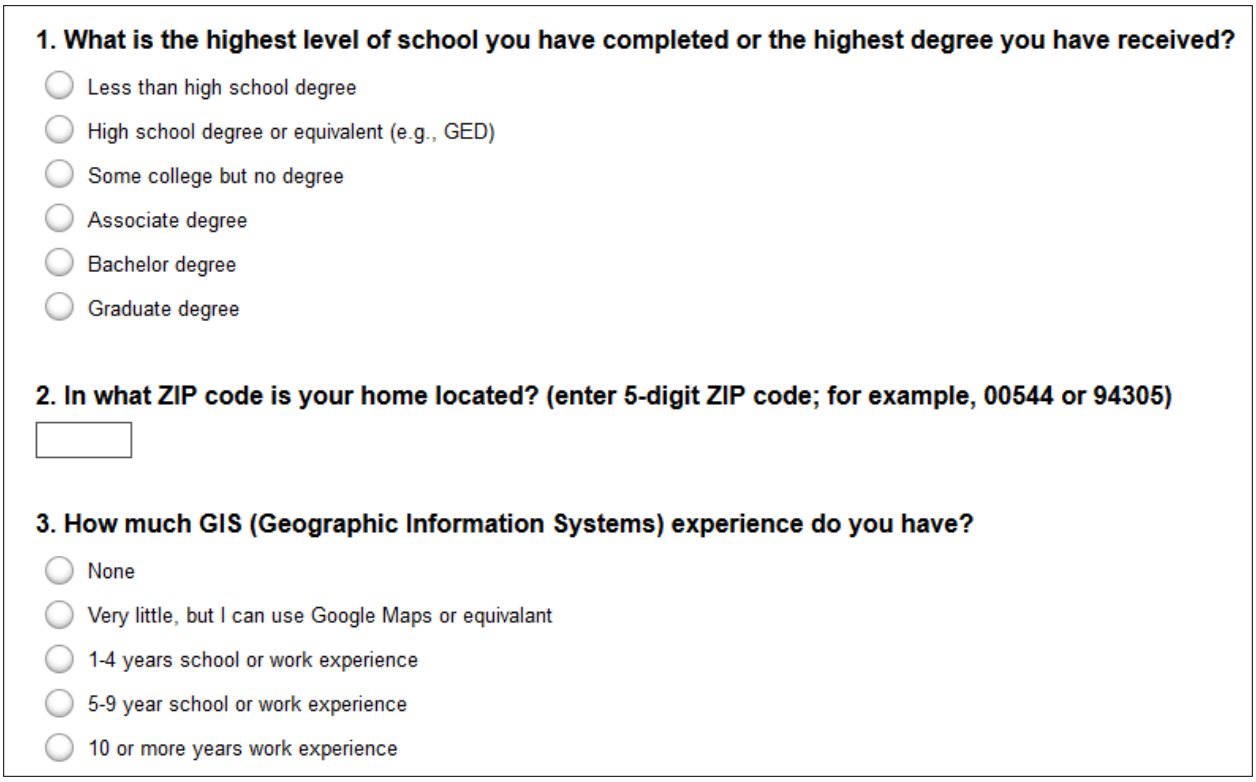
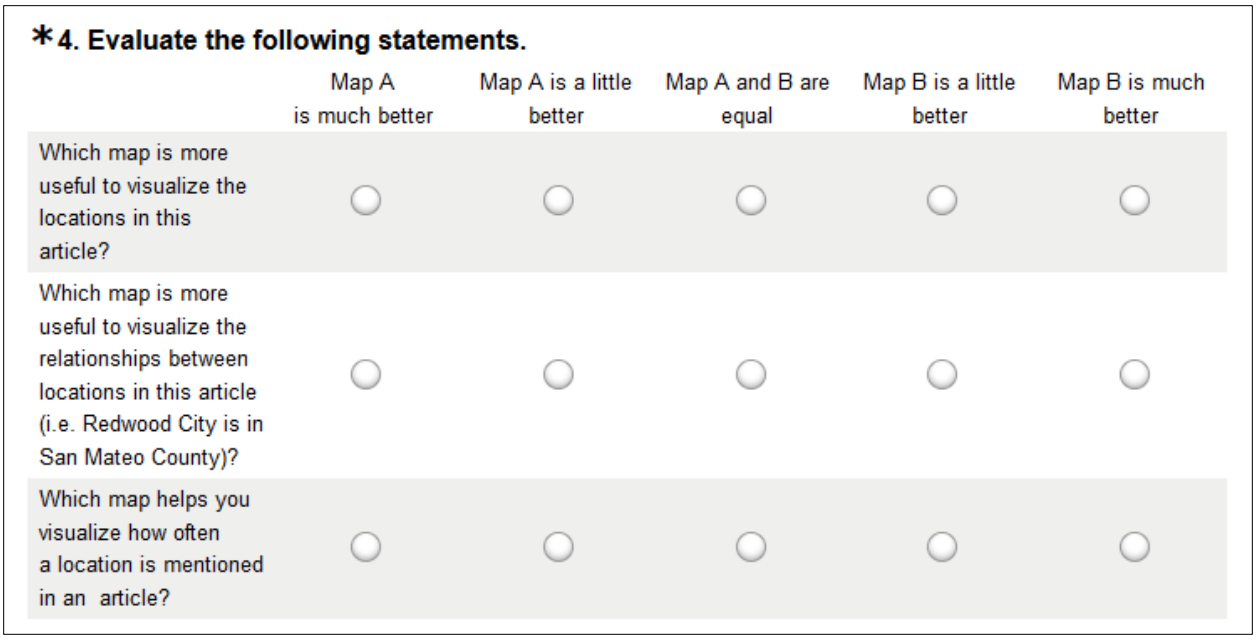


Figure 21 Survey Questions About the Respondents
Image by Survey Monkey (SurveyMonkey 2014)

In the main survey body, the user was presented with the news article at the top of the page and a screen print of MAP A, the SAV generated map with manually corrected PIG results for that article, as well as a screen print of MAP B, the CLAVIN map of the article. The respondents were then presented with three questions and the options to select which map was "a little better" or "much better" than the other, or if they were "equal;" an example is shown in Figure 22 below. The questions were used to find out which map the respondents thought helped them better visualize the locations, relationships, and how often a location occurred in the text. To control the responses to the questions, they were given five choices: map A is much better, map A is a little better, map A and B are equal, map B is a little better and map B is much better.



*4. Evaluate the following statements.

	Map A is much better	Map A is a little better	Map A and B are equal	Map B is a little better	Map B is much better
Which map is more useful to visualize the locations in this article?	○	○	○	○	○
Which map is more useful to visualize the relationships between locations in this article (i.e. Redwood City is in San Mateo County)?	○	○	○	○	○
Which map helps you visualize how often a location is mentioned in an article?	○	○	○	○	○

Figure 22 Survey Questions to Compare a Pair of Maps
Image by Survey Monkey (SurveyMonkey 2014)

After the comparison question, the respondents were asked if it was useful to see a map for this article and were encouraged to provide any additional comments they had, as seen in Figure 23 below. Asking the respondents if it was useful to have a map for the article was a question designed to find out if a text-to-map solution appeals to people when they are reading news articles. The comments were useful in the evaluation of the results because it provided transparency into what the respondent's logic was for some of his selections.

Chapter 1 describes the problem providing a quality text-to-map solution; there is a need for better automated visualization of text-to-map data than the existing point based maps. The survey's purpose was to determine whether others found the tool to be useful. This section will describe the aggregate results of the survey. Of the 270 possible respondents, thirty-six provided feedback. Of the thirty-six people who attempted the survey, twenty-two of the respondents completed it in its entirety, which was the basis for the survey result analysis, as incomplete results are not taken into account.

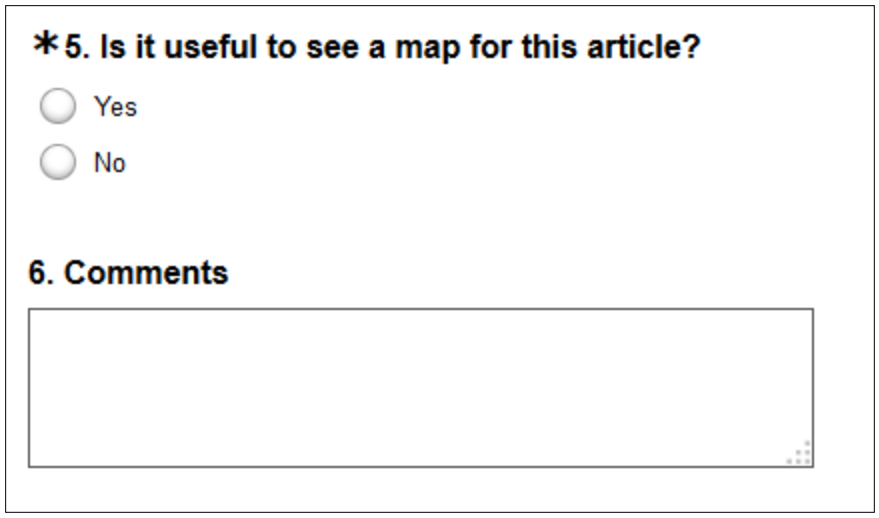


Figure 23 Survey Questions and Comments Format
Image by Survey Monkey (SurveyMonkey 2014)

4.4.2. Survey Results

The maps from a tool such as these would be targeted toward mainstream media such as e-book and online news publications. It is likely that the consumers of said maps would have little-to-no formal GIS experience. Therefore, it is important that the survey audience is representative of that demographic. The survey was not targeted at GIS students or professionals so that the value could be evaluated by those that represent the general population. When asked how much GIS experience the respondents have, the majority, 73%, responded that they have very little GIS experience. 9% or two respondents had 1-4 years school or work experience and 18% of respondents had no GIS experience at all.

In terms of education, all respondents have at least a high school degree and 77% have a bachelors or graduate degree, making it safe to assume the respondents have had at least a high school level exposure to geography.

It is important to note that fifteen of the twenty-two survey participants reside within the San Francisco Bay Area and seven live in other areas including: Southern California, Minnesota,

Illinois, Indiana and Texas. This indicates that the majority of respondents have some familiarity with the region; this is important to help understand if the maps are more useful for people who are familiar with the study area than those who are not.

Of the twenty-two survey respondents that were asked "Is it useful to see a map for this article?" 84% agreed that "yes," it was useful to have a map and 16% thought that "no," it was not useful, as seen in Figure 24. These numbers indicate that in general people are in favor of having maps with news articles. This question was asked for each article and the metrics were about the same, approximately 80% agreed that maps were useful.

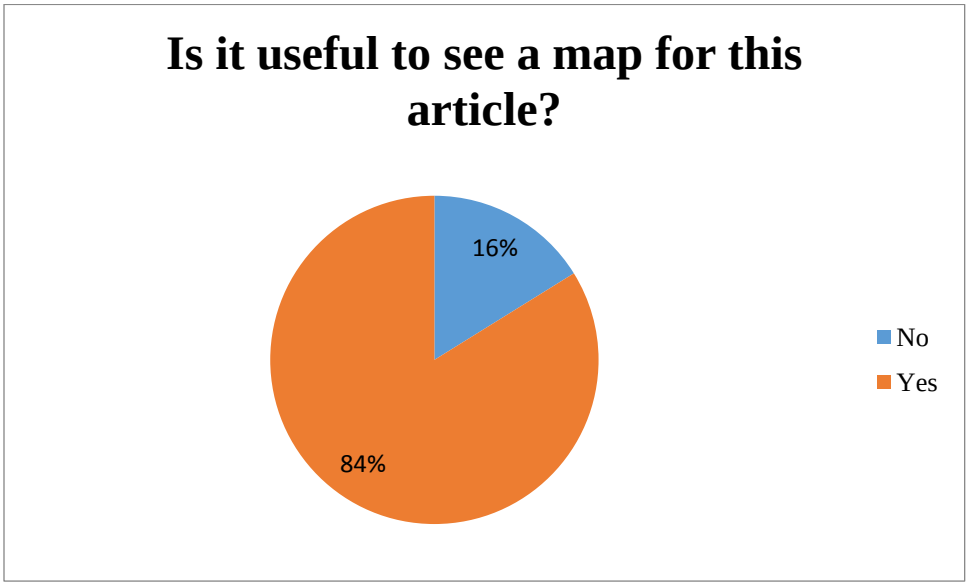


Figure 24 Graph of Survey Response Answers Aggregated for All Maps Image by Survey Monkey (SurveyMonkey 2014)

The article where the fewest respondents found a map useful, only 27%, was the article titled "Helix: A mini Exploratorium for the South Bay." This may be because the article is really about one specific location, the Helix Museum in Los Altos (Hicks 2014). The other locations mentioned in the article do not have much relevance to understanding the article. Conversely, the article titled "Pleasanton: New Iron Horse Trail segment opens"(Cuff 2014), 96% of respondents

felt it was useful to have a map with this article because the trail spans a larger area and understanding the relationship between the locations would help the reader visualize the new segment of trail.

When asked “Which map is more useful to visualize the locations in this article?” 74% of the time the response was in favor of the map generated by SAV, which was called Map A in this survey. 5% of the time they thought the SAV generated map and the CLAVIN online tool map were equal, and 22% of the time the CLAVIN online tool map was better; the breakdown can be seen in Figure 25.

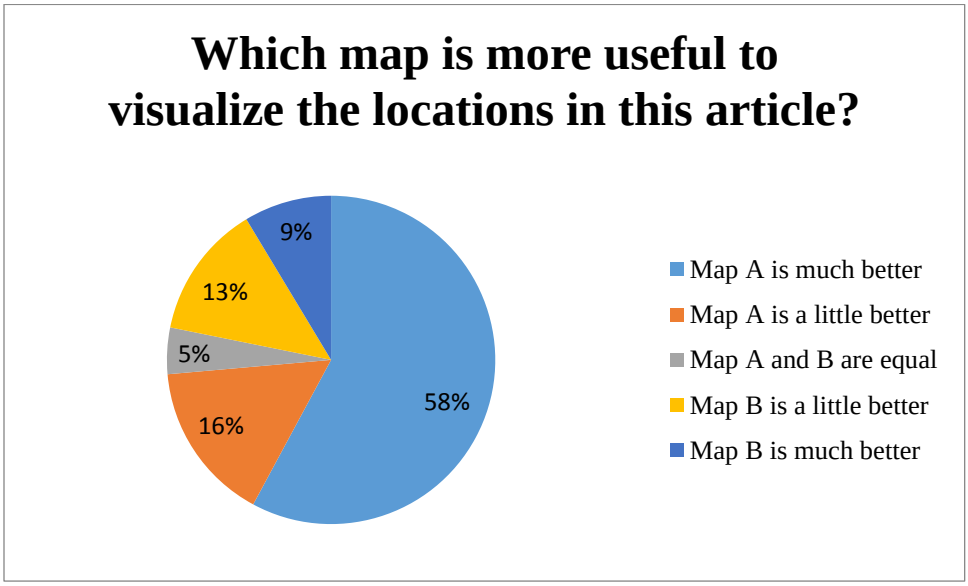


Figure 25 Aggregate Survey Response for Usefulness of Maps for Locations Image by Survey Monkey (SurveyMonkey 2014)

Next, the reponsdents were asked, “Which map is more useful to visualize the relationships between the locations in the articles?” 84% responded that the SAV generated map, Map A, was better at the visualization of the relationships. This is not surprising because the CLAVIN webtool generated map did not really specify any realtionships in the map; only 16%

found this map to be more useful for visualizing relationships. Figure 26 shows the aggregated results of all twenty-two respondents for all 9 maps.

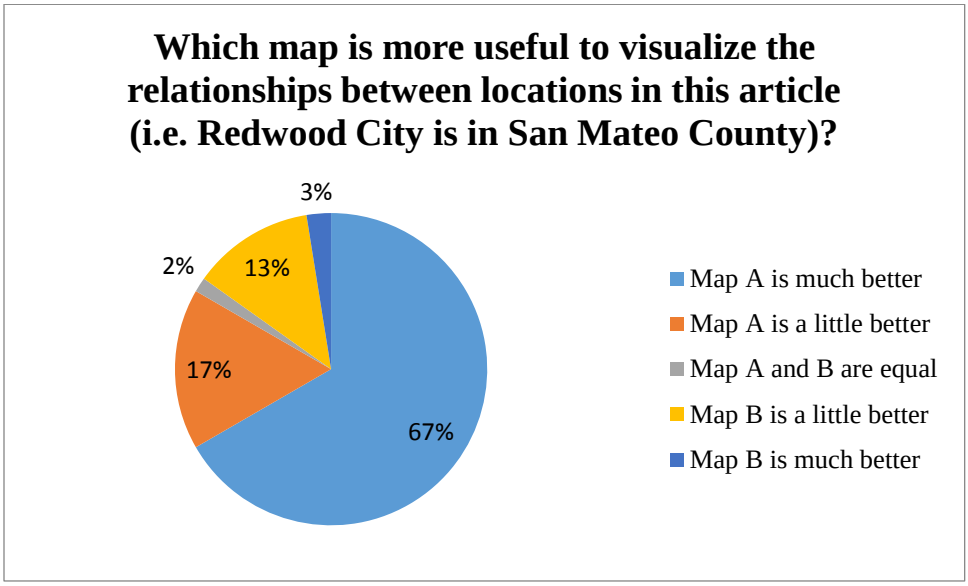


Figure 26 Aggregate Survey Response for Usefulness of Maps for Visualizing Relationships
Image by Survey Monkey (SurveyMonkey 2014)

The final question when comparing maps: "Which map helps you visualize how often a location is mentioned in an article?" 11% indicated a preference, for Map B, the CLAVIN online tool generated maps. This is surprising because those maps have no reference to how many times a location is mentioned in an article. The majority of the responses are still in favor of the SAV generated maps, as seen in Figure 27, with 82% of the responses finding the SAV generated map "a little" or "much" better.

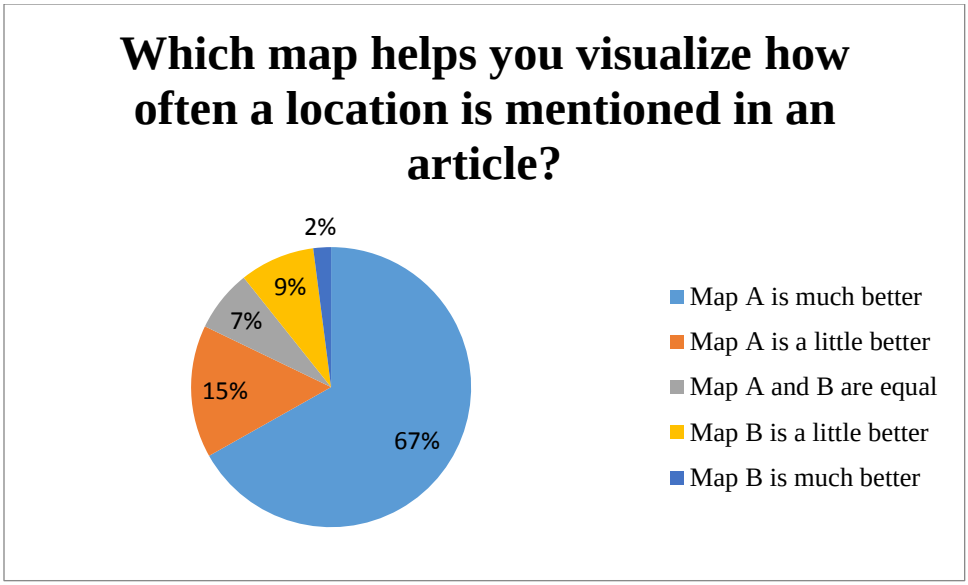


Figure 27 Aggregate Survey Response for Comparing Which Map is Better at Showing How Often a Location is Mentioned in the Text
Image by Survey Monkey (SurveyMonkey 2014)

The survey results showed that, in general, the respondents are in favor of the visualization that map A, the SAV generated map, provides. The results also clearly show that the way the relationships between locations were visualized by SAV was better than the existing CLAVIN online tool. The next section will provide a discussion on the results including the open-ended comments.

4.4.3. Survey Discussion

The survey results indicated that 84% of the time the respondents agreed, having a map to accompany an article was useful. In general, the SAV generated maps were well-received and the survey respondents found them useful. The majority of the time, the survey results indicated that the SAV generated map was preferred to the traditional point-based maps generated by the CLAVIN online tool. On the survey, each set of maps also had a free-form comments section to obtain the respondents' opinions on each set of maps. This free-form feedback was useful in

understanding why they may have chosen specific responses, as well as in collecting feedback for future improvements.

When asked “Which map is more useful to visualize the locations in this article” the majority, 74%, agreed that the SAV map was either a little or much better than the CLAVIN online tool map. One respondent found that the SAV generated map “is useful for people who don't know the area to get a "spatial feel" of the incident while/after reading the article.” Another respondent found that the CLAVIN online tool generated map was “slightly misleading and general” and another respondent commented that it was also “hard to relate to the article.” While most respondents preferred the SAV maps, one respondent noted that the SAV map “shows just the locations mentioned in the story, instead of all towns in the surrounding area” and that showing the locations in the surrounding area, like in the CLAVIN online tool map, would make it easier to understand where the locations are, which raises a good point. This is good feedback that could be incorporated into future versions.

By displaying the locations as polygons, the users are able to more clearly visualize the relationships between locations. 84% of respondents agreed that the SAV generated map was better at showing the these relationships. For the crime article about a shooting in San Ramon, a respondent found that the SAV generated map was particularly useful because it “showed the relationship between assailant’s home and location of shooting.” There were two comments that referred to the fact that the SAV map could be improved by displaying cities like San Francisco so people could understand the relationship between San Francisco and the locations on that map; this may explain why in some cases the CLAVIN generated map was selected as the better choice for displaying the relationship between locations. Overall, the SAV generated maps met

the objective of showing relationshipsbetween locations to help the user visualize the content of the article.

In response to the question “Which article helps you visualize how often a location is mentioned in the article,” 82% of the responses were in favor of the SAV map, which used categorization based on the frequency attribute to show color coded symbols for the number of times a locations was mentioned. However, in the free-form comments, there was an indication that for some of the articles, more specialized maps were desired rather than the symbology that was used to just highlight the locations in the text. One respondent commented in reference to an article about a disease outbreak that “(w)ith this article it would be helpful if the counties could be a darker shade if they have more cases.” Although this is a good suggestion, the intent of this map was not to show how many cases are in each location but merely where the location was mentioned in the text. There were two other cases where the respondents mentioned the need for a more specialized map. This comment does highlight the fact that although the majority of response data agreed having a map is useful, they may actually be looking for a more specific type of map that shows the actual outbreak data and not just the location in the text. In reference to the same article, another respondent stated that “This is the kind of article that stands to benefit greatly from simple mapping. Putting origins and recent reports of, say, a disease onto a map to show global-scale interactions.” This, again, points to the fact that for outbreak articles, a more specialized map may be required and an automated text-to mapping solution may not provide a useful final map in this situation. For the article about the Iron Horse trail construction project, it was also suggested that a project specific map would be more useful.

In instances like the outbreak article and the Iron Horse trail article where just mapping the locations mentioned in the article is not as useful as a specially designed map, the SAV tool

could be used by someone as a starting point to create a map from the locations in the text and then they could manually adjust the map to make sure that the locations and symbology are meaningful to the map. SAV would be a useful tool to accelerate the workflow of map creation for news articles but may not always generate the final result.

SAV performed well in the mapping of locations from text with the exception of two distinct cases where better quality input data could resolve the issues. The survey responses indicated that the SAV generated maps were useful and performed better that the CLAVIN online tool maps. The next chapter will summarize the thesis project and discuss opportunities for future work.

CHAPTER FIVE: CONCLUSION AND FUTURE WORK

This chapter discusses how the challenges identified in Chapter 1 are addressed using SAV, as well as its limitations, providing an opportunity to discuss future work.

5.1. Results of SAV

The problem described in Chapter 1 is that there is currently no robust way, known to the author, to automatically extract spatial information from unstructured text and present it on a map displaying locations' boundaries and spatial relationships between locations. Chapter 2 reviewed existing work on NLP and NER techniques that are used by software to extract locations from text, as well as some online tools that perform a text-to-map function.

Chapter 3 explained the methods used by this thesis project to overcome the challenges described in Chapter 1. First CLAVIN, an existing geoparsing software was used to extract the spatial entities and geo-lookup the results. Second, the DIP toolset was developed to integrate the data and create relationships between the locations. Finally, the challenge of automatically creating a map was addressed by using templates for the map layout, symbology and label styles in order to simplify the map creation process.

Chapter 4 explained how the results were evaluated using precision and recall metrics and survey responses. These metrics showed that that the NER portion of the software had challenges requiring manual editing of the PIG results. The survey results indicated that people find it useful to have a map along with a news article. The results also indicated that a map showing the relationships between the locations was, in general, a more useful tool for visualization than those that do not show the relationships; this indicated there is a need for more work on solutions like SAV.

5.2. Future work and Limitations

The prototype application (SAV) developed as part of this thesis project achieved its goal of providing a better way to visualize locations in text on a map. However, as with any project, there is always room for improvement. There are five areas where the thesis project could be enhanced: full automation, user interface, NER, Data Quality, and geographic scope. With these enhancements this approach could be incorporated for use with news agencies or e-book publishers to quickly generate maps from text.

One of the easiest opportunities for future work would be to fully automate the approach so that the user can just enter an article's URL and the program would identify the article name and article text to extract. Automating the article capture, using a website scrapping software, would eliminate the need for the user to manually create and maintain separate article files and the table with article name and URL. The next logical step would be to fully automate the flow between the Java class and the DIP toolset created in ArcMap. One limiting factor of the fully automated approach is that the results from PIG are not always accurate enough to process directly; in a future application, one would need to allow a user to override the locations extracted by PIG.

Since the current program is just a prototype the user interface is not very user friendly. The user must manually prepare the article text and execute the java and ArcGIS processes separately. To make the program useable for preparing maps for news articles or e-books a user interface would need to be developed, likely a web interface, to allow the users to interact with a single system for all the processing steps.

The NER portion of the thesis project is probably the area that requires the most work. As discussed in Chapter 2, in order for machines to be able to parse text for locations and understand

which of the ambiguous locations corresponds to the text, more research is required. As NLP and computer learning techniques improve, the visualization portions of this thesis project could be adapted to work with different NER software. The approach used by this project to integrate with the polygon boundaries could be used to help resolve location ambiguity in the NER results by testing if locations in a document have relationships such as adjacency or containment. The polygon boundaries would need to be integrated with the point data earlier in the process where the disambiguation occurs.

Data quality seemed to provide the biggest challenge. Selecting a different data set could improve the data quality in the future. The San Francisco Bay Area polygon file that was selected did not have the enough detail in its attribute data. If the polygon file had used the same Feature Class and Feature Code classification as the point features, it would have been easier to determine which locations were matches. Since the data sets are provided by organizations that are open source and promote community involvement in maintaining the data, issues with the data could be reported to the organizations as they are recognized. In this way, the organization can correct the data problem and the quality of the existing data set can improve. The tools used to join and filter the points and polygons could be used to join the Geonames points and Metro Extract polygons and then perform data analysis to identify inconsistencies in the data sets.

Limiting the scope of the experiment may have slightly oversimplified the thesis project. If a state, county, or global geographic scope was used, then there would be more challenges to select the correct extent for the map display to be able to see the necessary details. The filtering of boundary locations would be more challenging if there were more administrative levels present. If the solution was expanded to a larger area to include more levels, such as state and country boundaries, the logic would have to be adjusted in two major areas. The first area for

improvement would be to understand how many levels of the administrative boundary hierarchy the users find useful. This would be important for expanding the logic used in the thesis project. The second area where the logic would need to be adjusted is the scale for the main map. With a global scope, more work would need to be done on the logic for setting up multiple data frames and deciding which locations should go into each data frame. In this experiment, since the study area was known, it was easy to separate the data into the correct data frames.

Finally, this thesis project demonstrates that the known existing map-to-text applications do not provide any logic around the relationships between locations and visualization, other than plotting points on the map. The case study provided an example of the method used by SAV to create relationships and visualization of locations. The evaluation of survey results shows that SAV was effective in generating a better map than the existing solutions. The comments on the survey results showed that although maps are useful with articles, there is work that can be done to improve the SAV results. Respondents were interested in maps that were more contextual to the situation. There is work to be done in the NLP area to be able to automate the process to develop maps like this. In the meantime, SAV could be a useful process for people to use to generate a map to start with and edit based on the context of the locations in the articles.

REFERENCES

2014a. "About GeoNames." GeoNames Accessed 8 May 2014. http://www.geonames.org/about.html.

2014b. "OpenCalais Documentation." Thomson Reuters Accessed 26 Sep 2014. http://www.opencalais.com/documentation/opencalais-documentation.

2014c. "PlaceSpotter - YDN." Yahoo! Inc. Accessed 26 Sep 2014. https://developer.yahoo.com/boss/geo/docs/key-concepts.html.

2014d. "Readme for GeoNames Gazetteer extract files." GeoNames Accessed 8 May 2014. http://download.geonames.org/export/dump/readme.txt.

Amitay, Einat, Nadav Har'El, Ron Sivan, and Aya Soffer. 2004. "Web-a-where: Geotagging Web Content." Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY.

Berico Technologies, LLC. 2014. "CLAVIN Web Application." Accessed 6 Aug 2014. http://clavin.berico.us/clavin-web/.

Cambria, Erik, and Bebo White. 2014. "Jumping NLP curves: A review of natural language processing research." *IEEE Computational Intelligence Magazine* 9 (2):48-57.

Craft, Cynthia H. 2014. "More whooping cough cases reported by California, counties." Accessed 11 May 2014. http://www.sacbee.com/news/local/health-and-medicine/article2589126.html.

Cuff, Dennis. 2014. "Pleasanton: New Iron Horse Trail segment opens." Accessed 4 Sep 2014. http://www.insidebayarea.com/livermore/ci_26409559/pleasanton-new-iron-horse-trail-segment-opens.

Dodge, Martin, Mary McDerby, and Martin Turner. 2008. "The power of geographical visualizations." *Geographic Visualization: Concepts, tools, and applications*:1.

D'Ignazio, Catherine, Rahul Bhargava, Ethan Zuckerman, and Luisa Beck. 2014. "CLIFF-CLAVIN: Determining Geographic Focus for News."

Esri. 2014a. "ArcGIS What's New in ArcGIS." Accessed 5 Oct 2013. http://www.esri.com/software/arcgis/arcgis10/.

Esri. 2014b. "Light Gray Canvas." esri Accessed 5 Oct 2014. http://www.arcgis.com/home/item.html?id=8b3d38c0819547faa83f7b7aca80bd76.

Gelernter, Judith, and Wei Zhang. 2013. "Cross-lingual geo-parsing for non-structured data." Proceedings of the 7th Workshop on Geographic Information Retrieval, 11/05/2013.

Gomez, Mark. 2014. "San Jose: 16-year-old charged as adult in July homicide." Accessed 26 Aug 2014. http://www.mercurynews.com/crime-courts/ci_26387583/san-jose-16-year-old-arrested-and-charged.

Handa, Robert. 2014. "Hate Crime Charge in Palo Alto Beating Case." Accessed 3 Sep 2014. http://www.nbcbayarea.com/news/local/Palo-Alto-Hate-Crime-Charge-in-Beating-Case-272939951.html.

Hicks, Tony. 2014. "Helix: A mini Exploratorium for the South Bay." Accessed 5 Oct 2014. http://www.mercurynews.com/travel/ci_24952630/helix-mini-exploratorium-south-bay.

Jewett, Daniel. 2014. "Greenbrae - Marin Magazine - February 2009 - Marin County, California." Accessed 10/24. http://www.marinmagazine.com/February-2009/Greenbrae/.

Johnson, Autumn. 2014a. "51 Confirmed Measles Cases in California Prompts Officials to Issue Warning." Accessed 8 May 2014. http://sanramon.patch.com/groups/politics-and-elections/p/51-confirmed-measles-cases-in-california-prompts-officials-to-issue-warning.

Johnson, Autumn. 2014b. "Family of UC Berkeley Student With Measles Contracts the Disease." Accessed 8 May 2014. http://patch.com/california/berkeley/family-of-uc-berkeley-student-with-measles-contracts-the-disease.

Johnson, Autumn. 2014c. "San Ramon Man Arrested After Allegedly Firing into Ex-Girlfriend's House." Accessed 8 May 2014. http://patch.com/california/sanramon/san-ramon-man-arrested-after-allegedly-firing-into-exgirlfriends-house.

Keller, Mikaela, Clark C Freifeld, and John S Brownstein. 2009. "Automated vocabulary discovery for geo-parsing online epidemic intelligence." *BMC Bioinformatics* 10 (1):385. doi: info:pmid/19930702.

Kelm, Pascal, Sebastian Schmiedeke, and Thomas Sikora. 2011. "Multi-modal, multi-resource methods for placing flickr videos on the map." Proceedings of the 1st ACM International Conference on Multimedia Retrieval.

Lieberman, Michael D., Hanan Samet, Jagan Sankaranarayanan, and Jon Sperling. 2007. "STEWARD: Architecture of a Spatio-textual Search Engine." Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems, New York, NY.

Lieberman, Michael D., Hanan Samet, and Jagan Sankaranayananan. 2010. "Geotagging: Using Proximity, Sibling, and Prominence Clues to Understand Comma Groups." Proceedings of the 6th Workshop on Geographic Information Retrieval, New York, NY.

McCurley, Kevin S. 2001. "Geospatial Mapping and Navigation of the Web." Proceedings of the 10th International Conference on World Wide Web, New York, NY.

McMaster, Robert B. Howard Hugh H. Kessler Fritz C. Slocum Terry A. 2005. *Thematic cartography and geographic visualization*. Upper Saddle River, N.J.: Pearson/Prentice Hall.

Migurski, Michal. 2014. "Metro Extracts." Accessed 8 May 2014. http://metro.teczno.com/.

Nadeau, David, and Satoshi Sekine. 2007. "A survey of named entity recognition and classification." *Lingvisticae Investigationes* 30 (1):3-26.

Nikolajevs, Jurijs, and Gints Jekabsons. 2013. "Automatic extraction of geographic context from textual data." *Applied Information and Communication Technologies (Latvia)*.

Rauch, Erik, Michael Bukatin, and Kenneth Baker. 2003. "A Confidence-based Framework for Disambiguating Geographic Terms." Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References - Volume 1, Stroudsburg, PA.

Strayed, Cheryl. 2013. *Wild: From Lost to Found on the Pacific Crest Trail*: Random House LLC.

SurveyMonkey. 2014. "Everything You Need to Know About SurveyMonkey." Accessed 5 Oct 2014. https://www.surveymonkey.com/mp/aboutus/.

Yadegaran, Jessica, and Linda Zavoral. 2014. "Oakland chef wins Esquire's TV's 'Knife Fight'; Mixt comes to the South Bay; Jack's Oyster Bar opens in Oakland." Accessed 8/26. http://www.mercurynews.com/eat-drink-play/ci_26381340/oakland-chef-wins-esquires-tvs-knife-fight-mixt.

APPENDIX A: MAP EXPERIMENT RESULTS

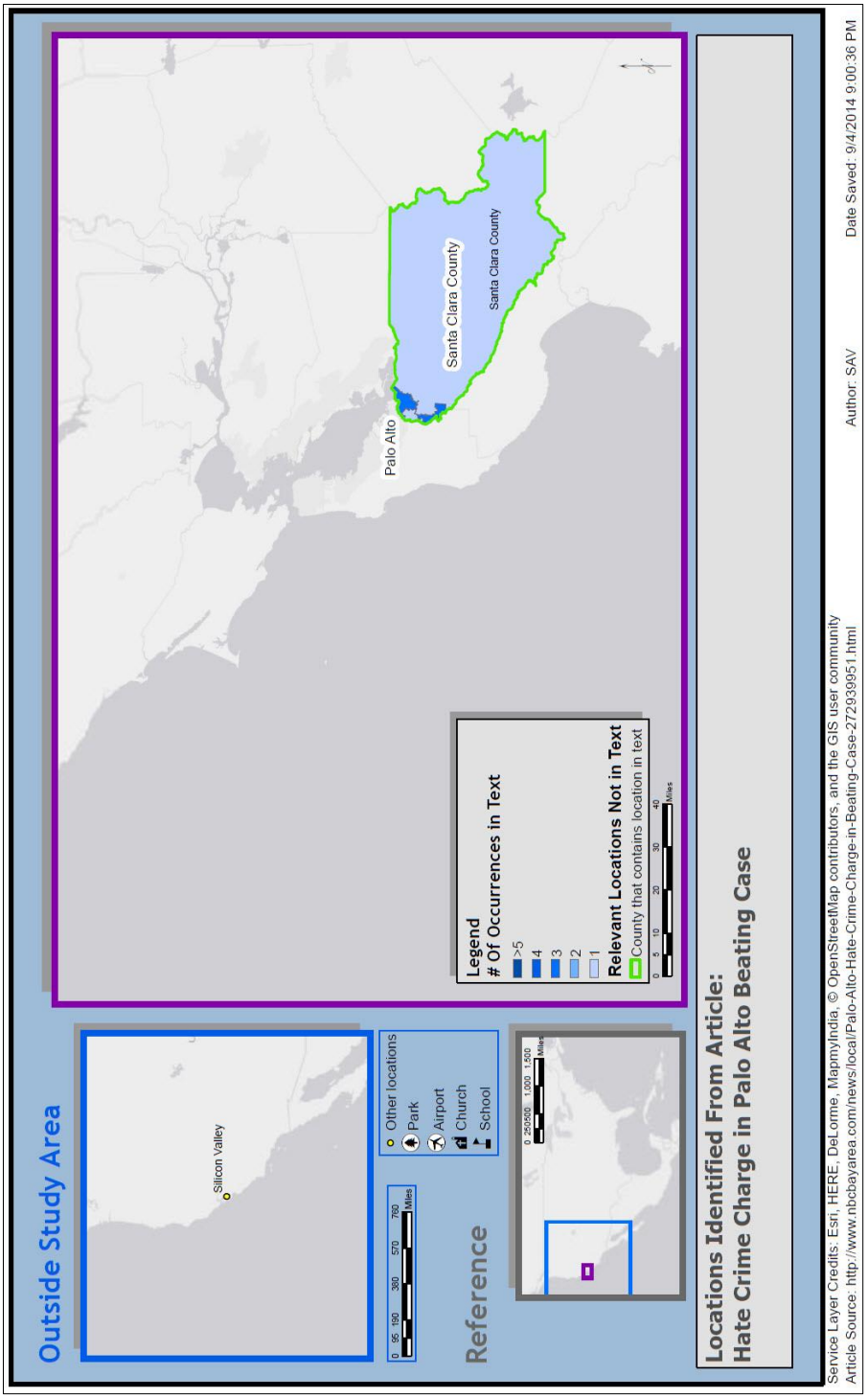


Figure A-1 SAV Generated Map for Hate Crime Article

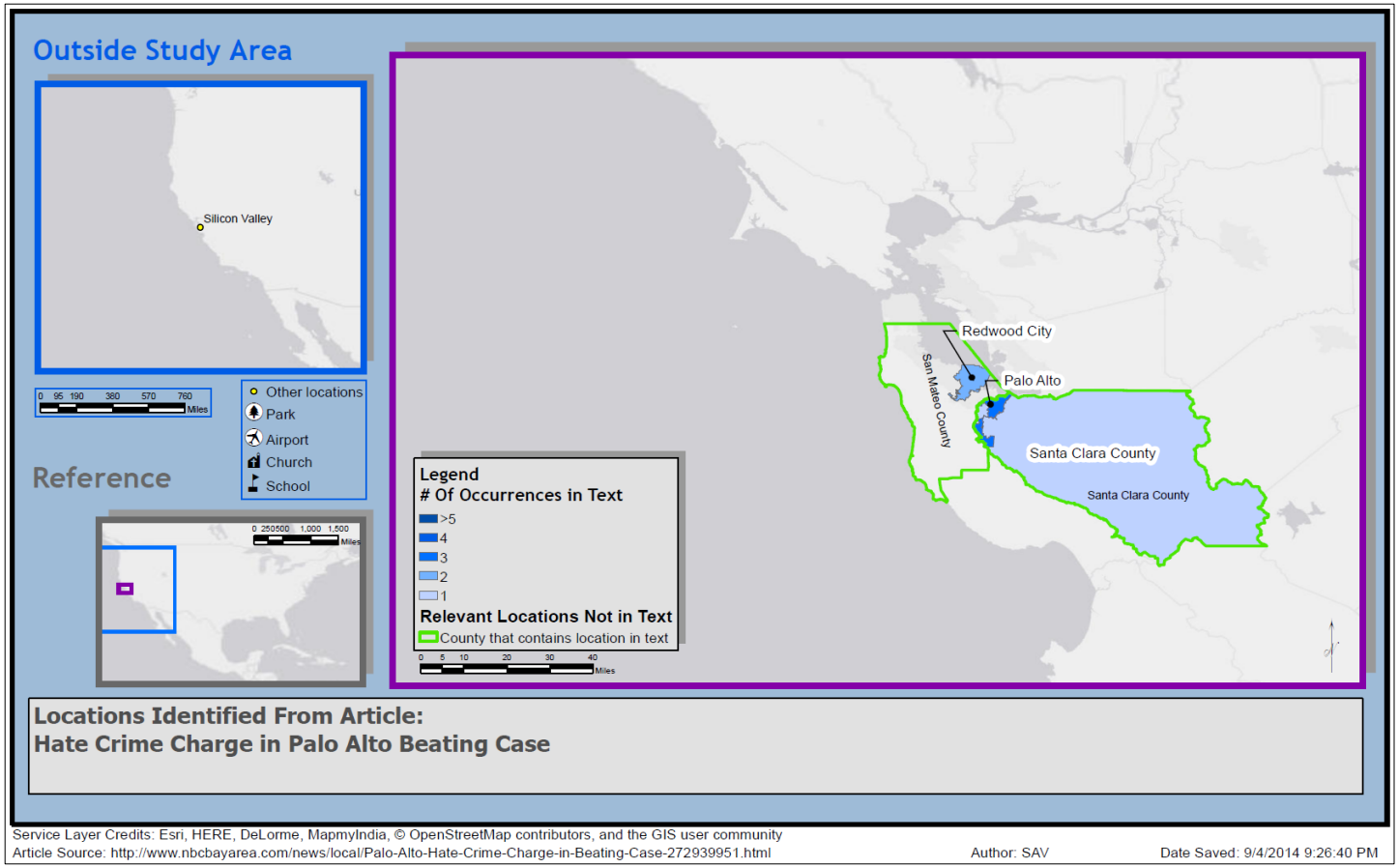


Figure A-2 Map with Manually Corrected PIG Results for Hate Crime Article

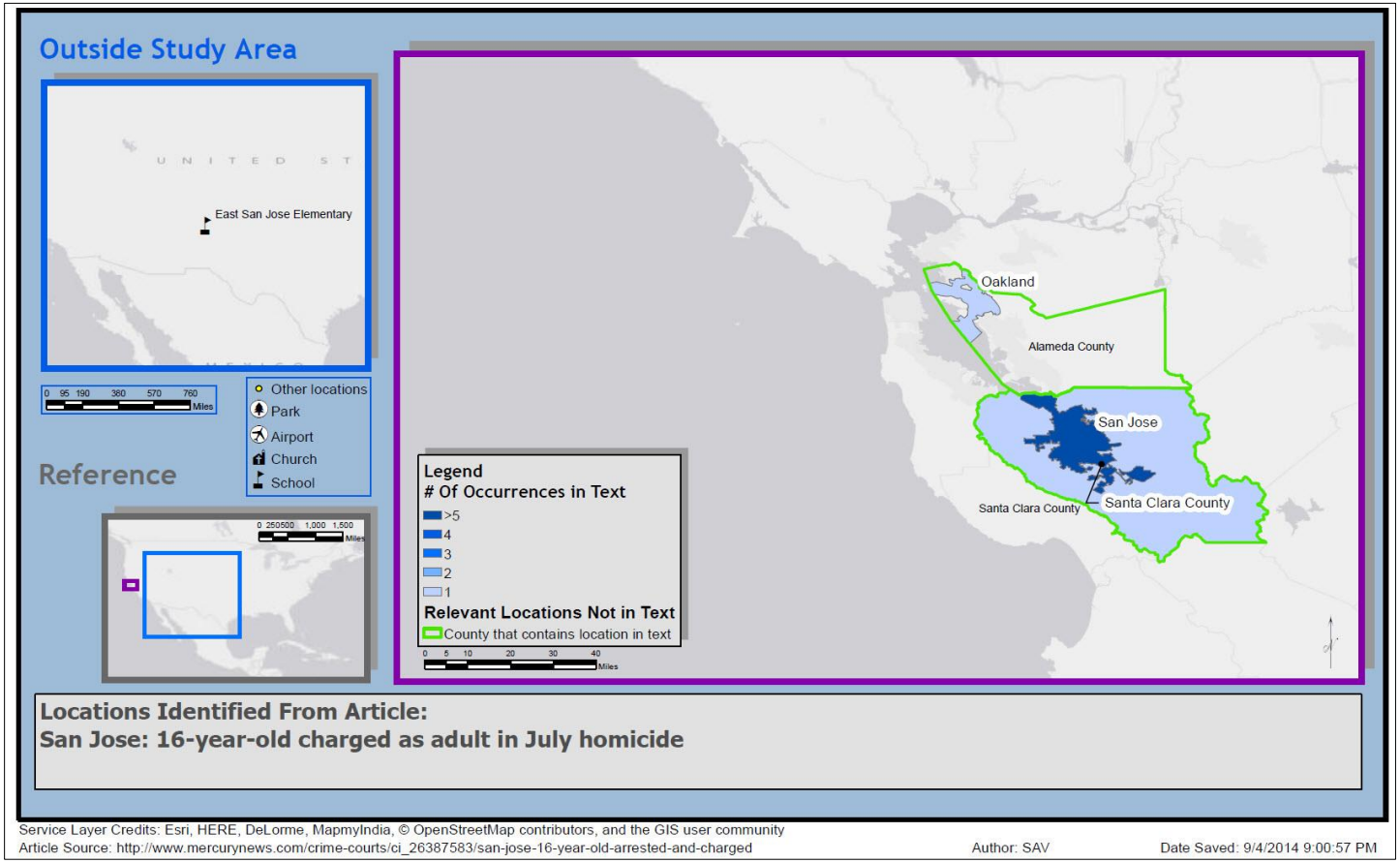


Figure A-3 SAV Generated Map for Homicide Article

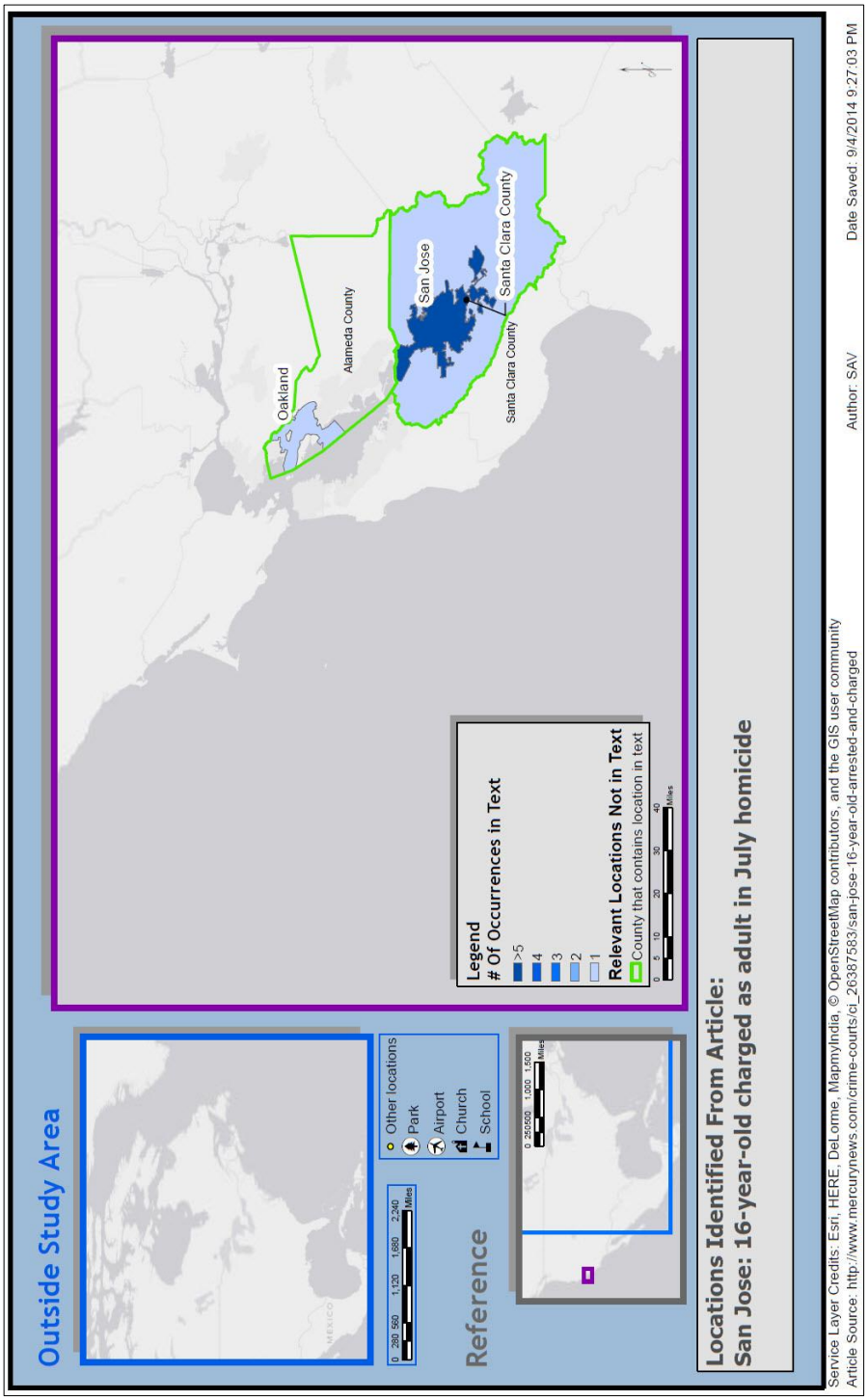


Figure A-4 SAV Generated Map with Manually Corrected PIG Results for Homicide Article

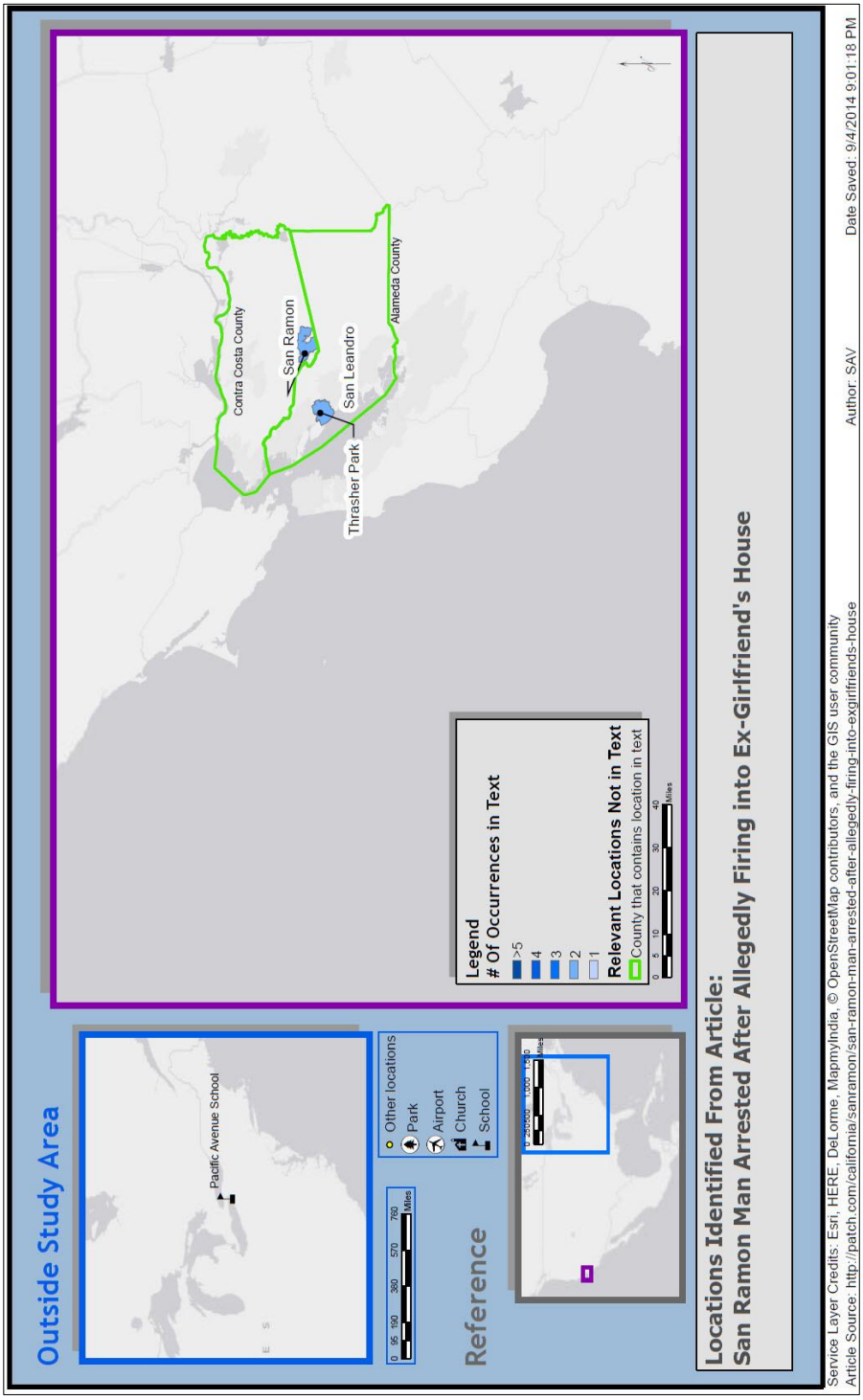


Figure A-5 SAV Generated Map for Shooting Article

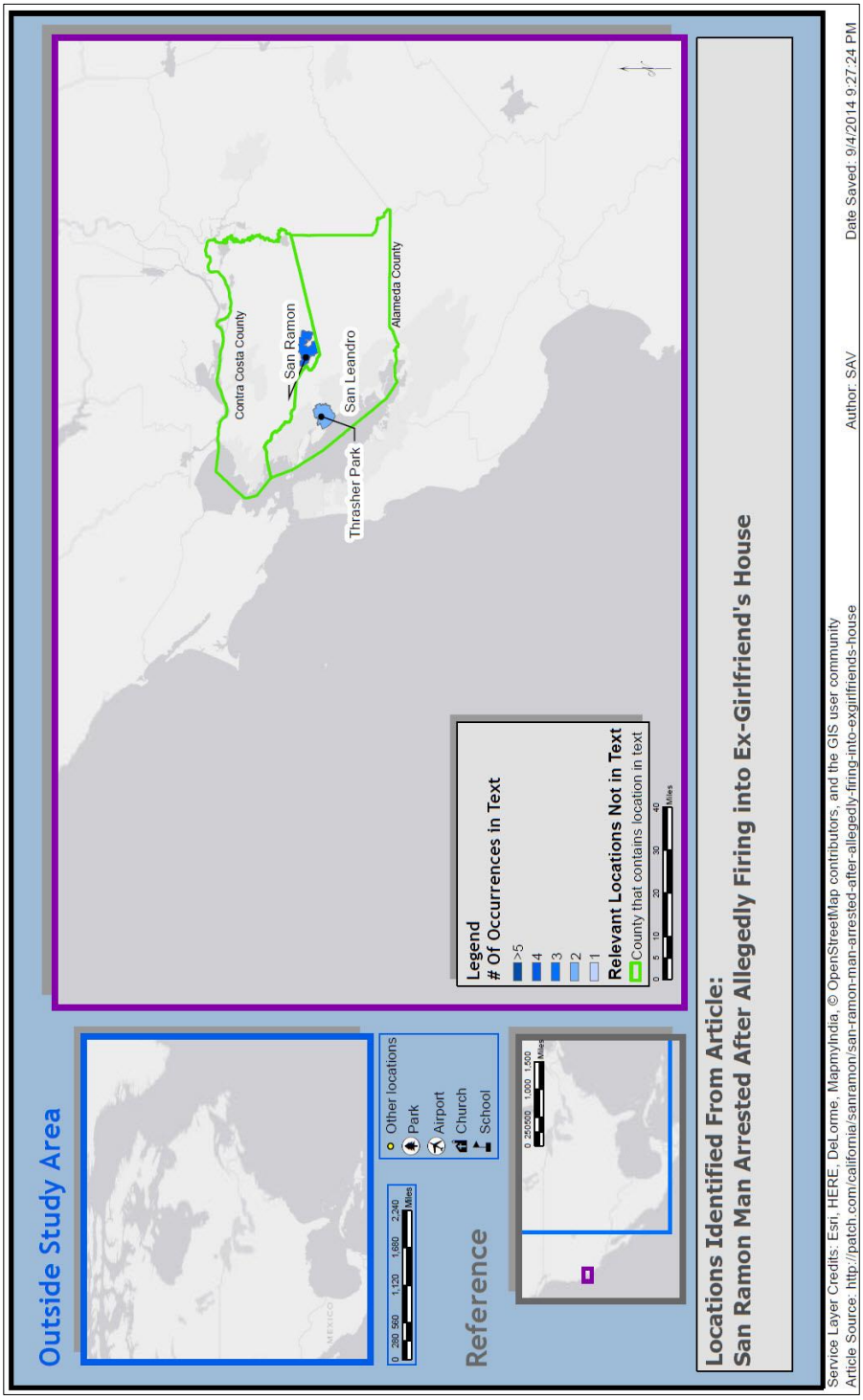


Figure A-6 SAV Generated Map with Manually Corrected PIG Results for Shooting

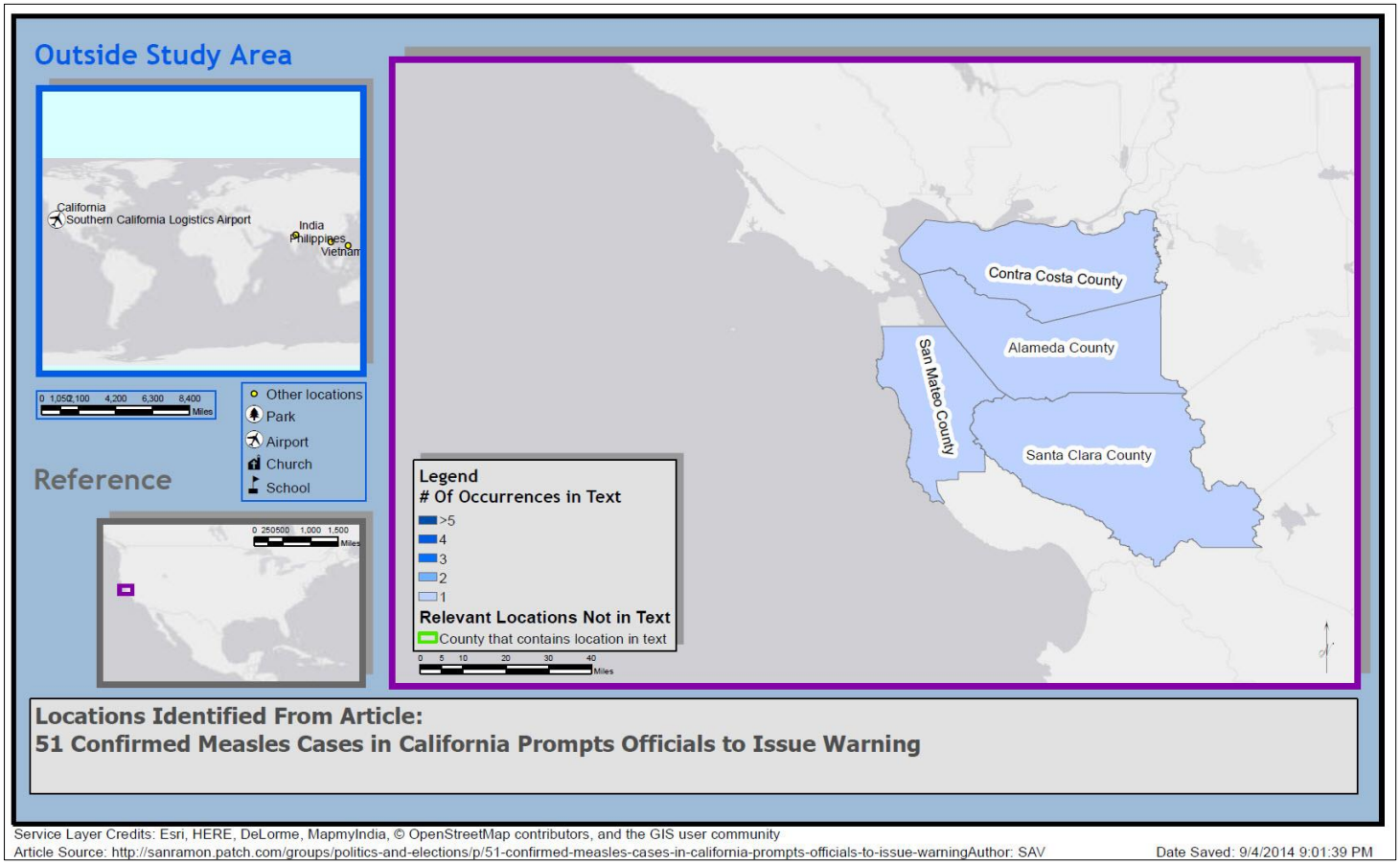


Figure A-7 SAV Generated Map for Measles Article

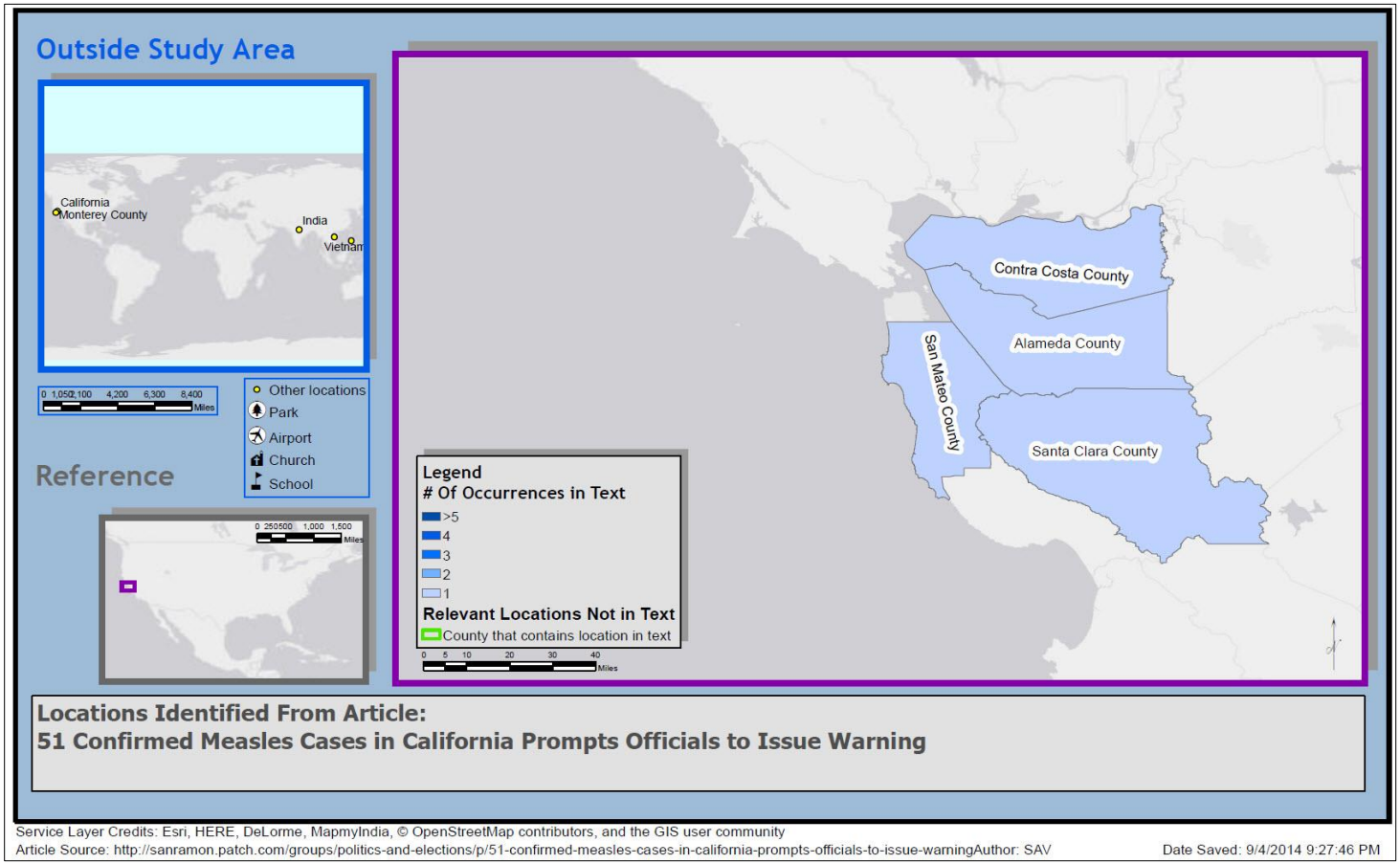


Figure A-8 SAV Generated Map with Manually Corrected PIG Results for Measles Article

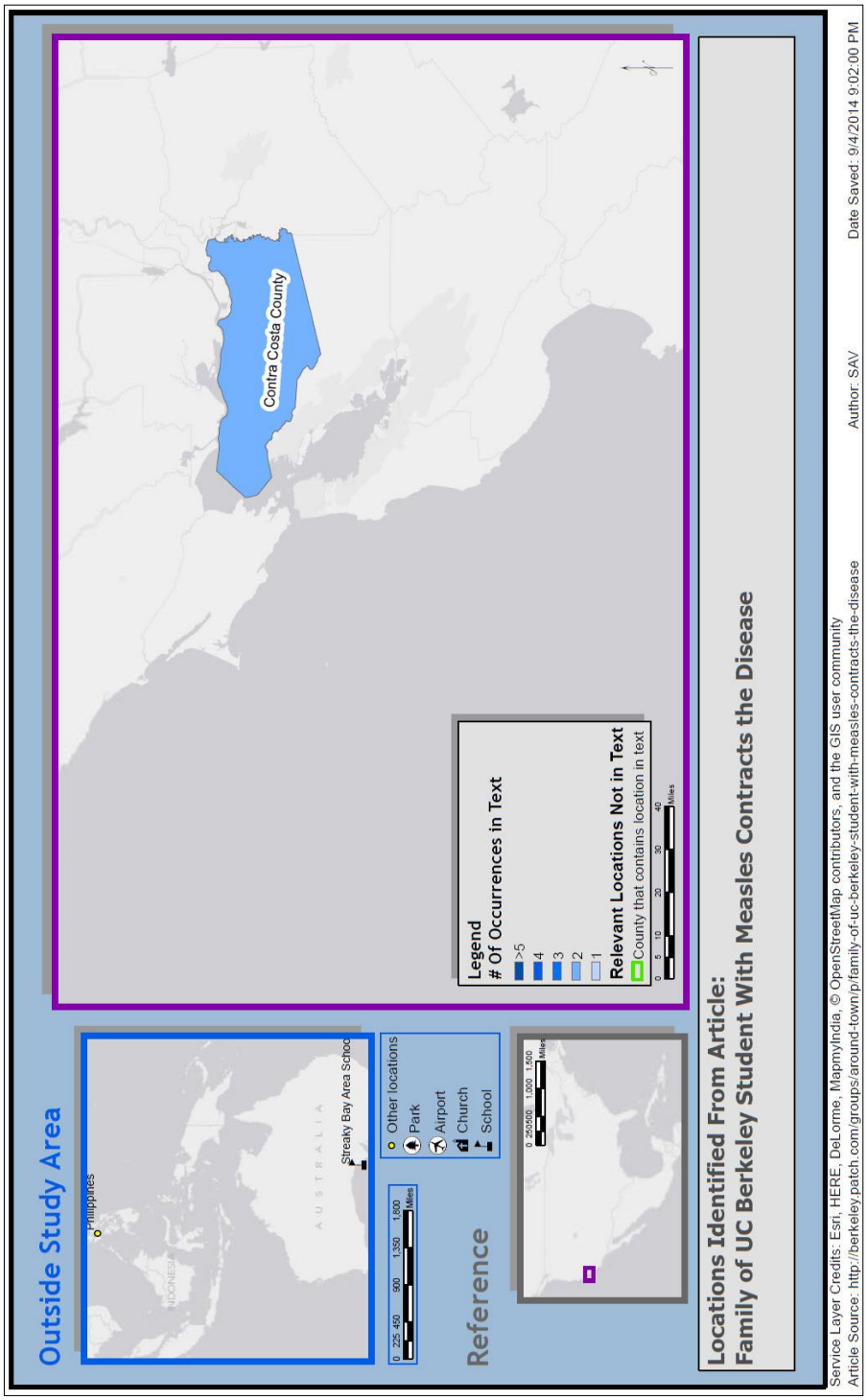


Figure A-9 SAV Generated Map for Measles Berkeley Article

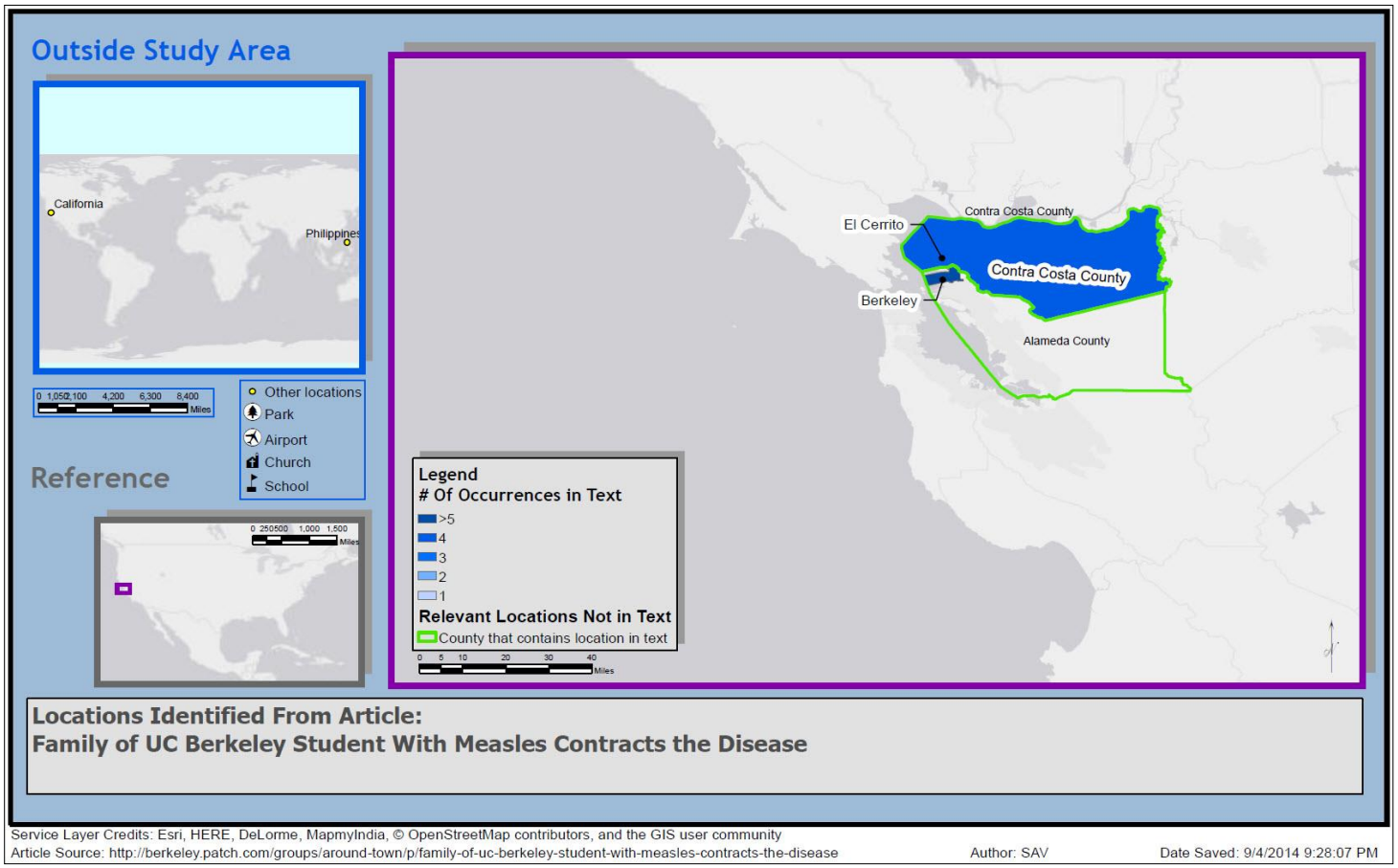


Figure A-10 Map with Manually Corrected PIG Results for Measles Berkeley Article

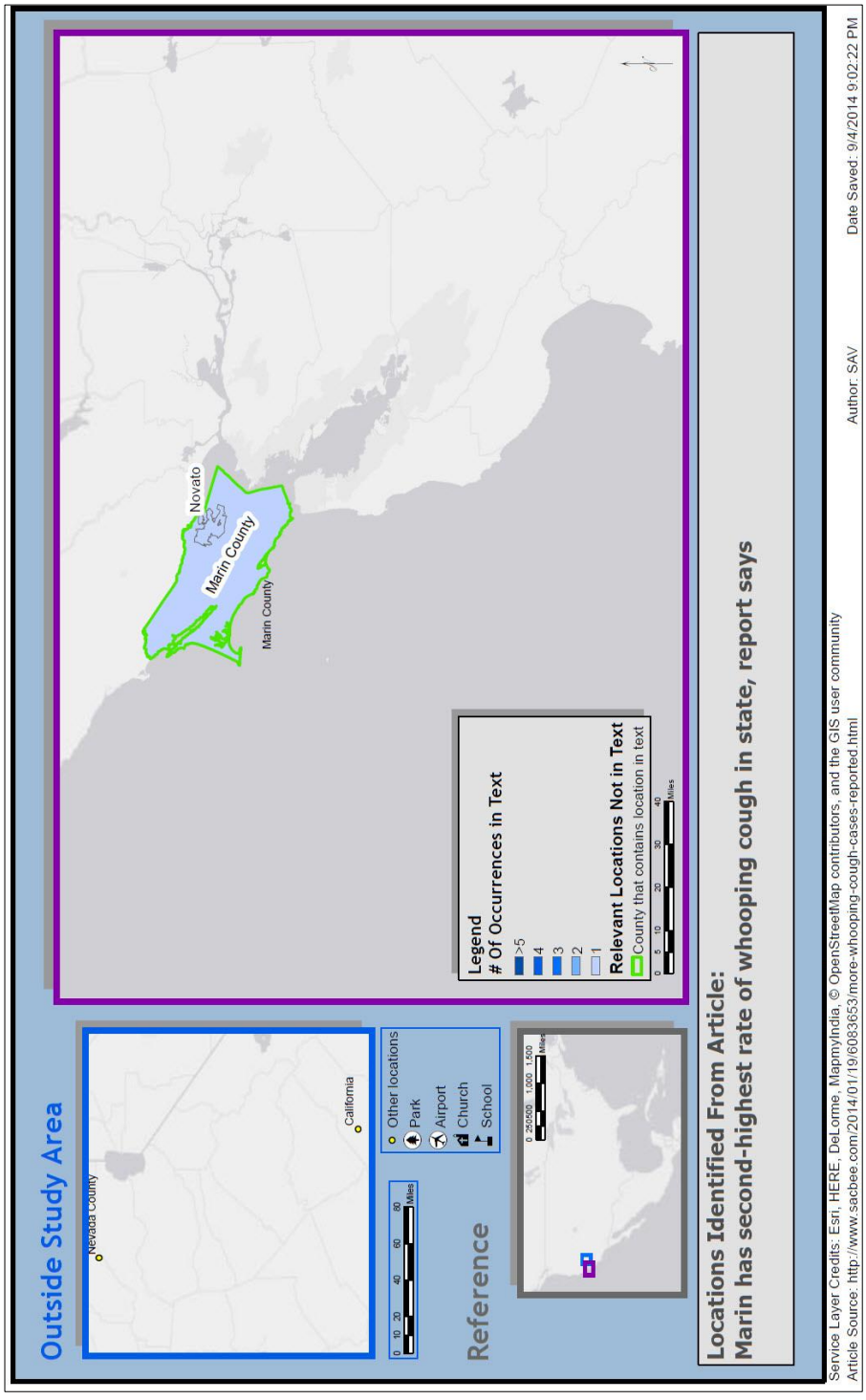


Figure A-11 SAV Results Whooping Cough Article Map

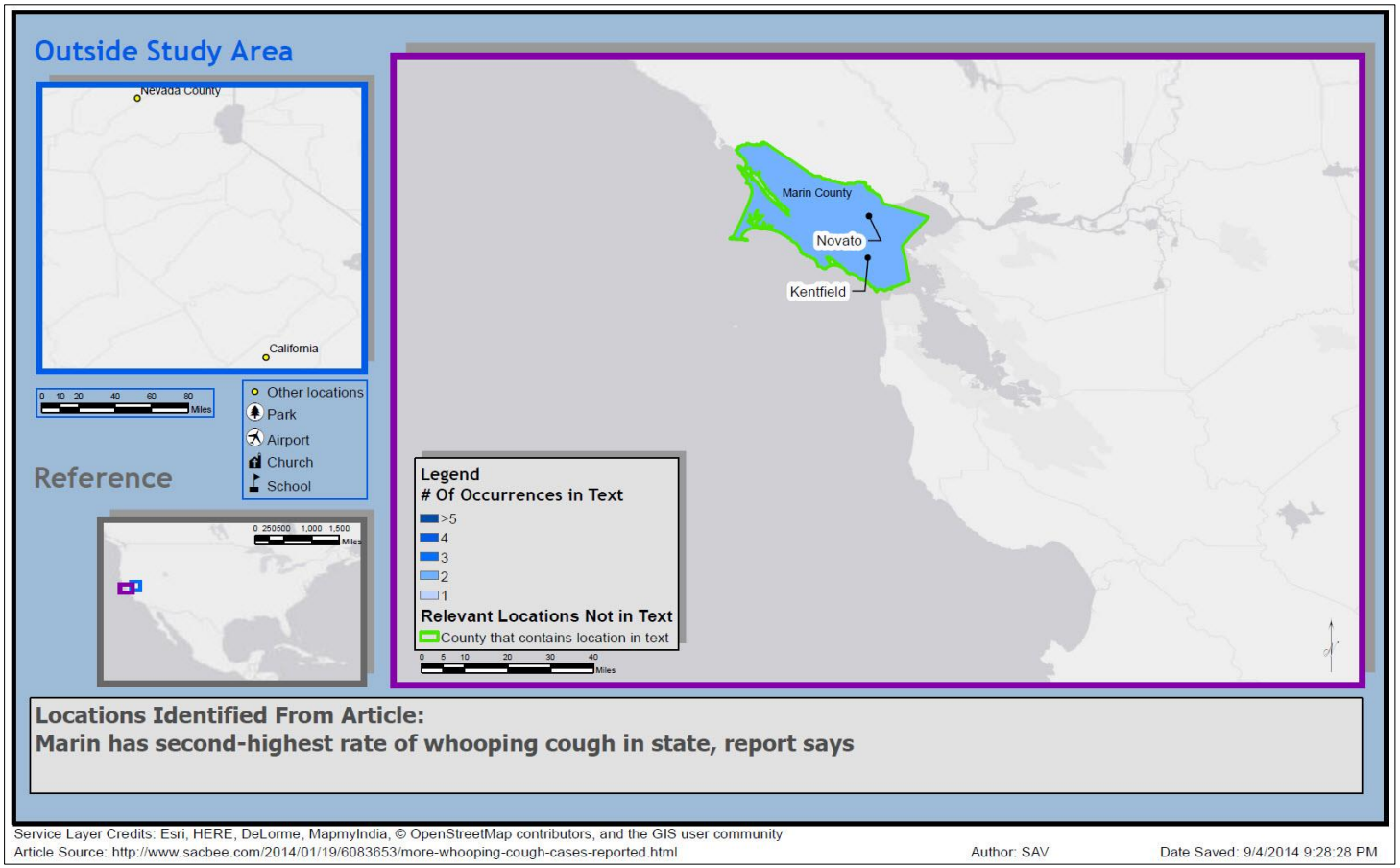


Figure A-12 Map with Manually Corrected PIG Results Whooping Cough Article Map

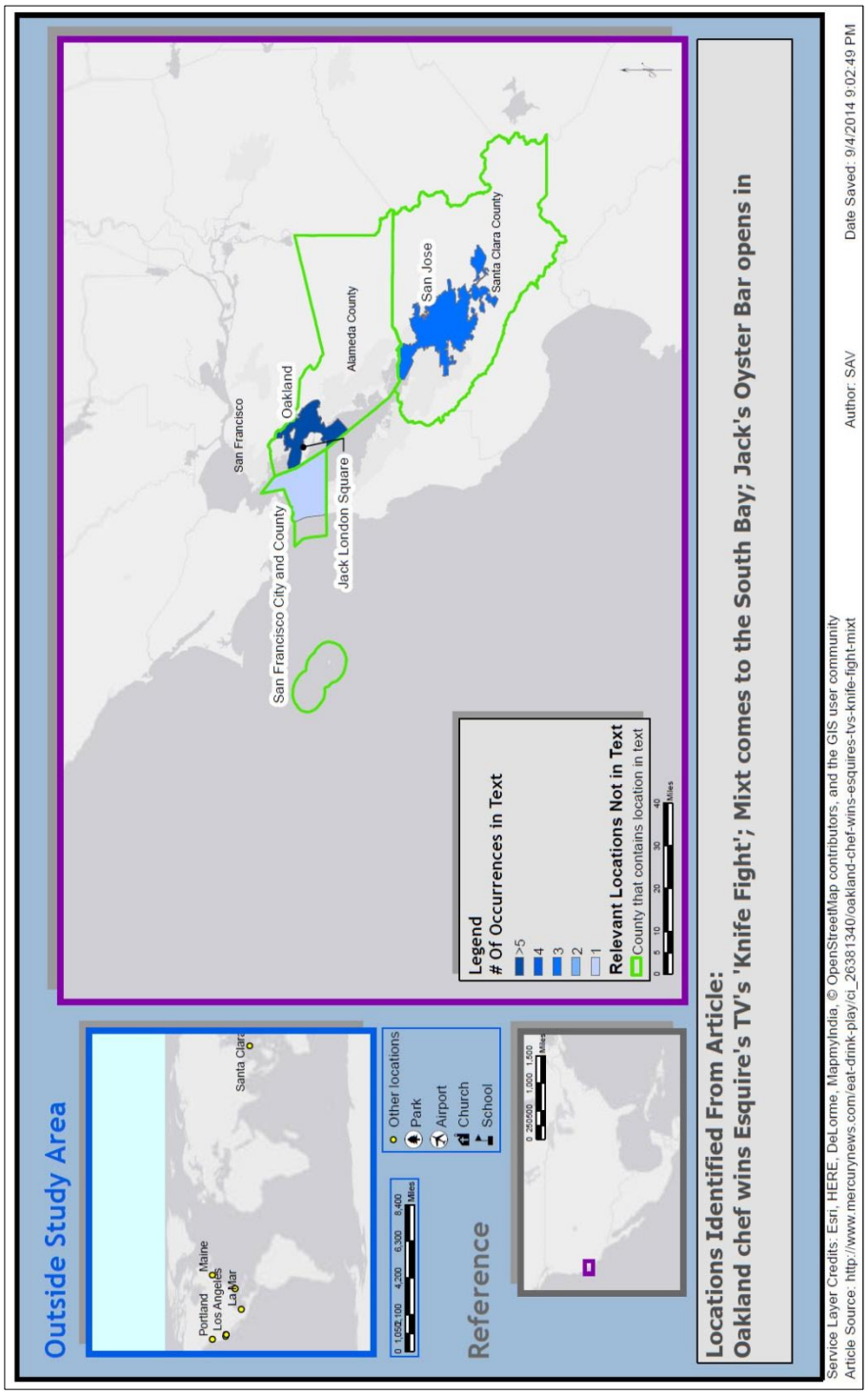


Figure A-13 SAV Generated PIGs Results for Food Article

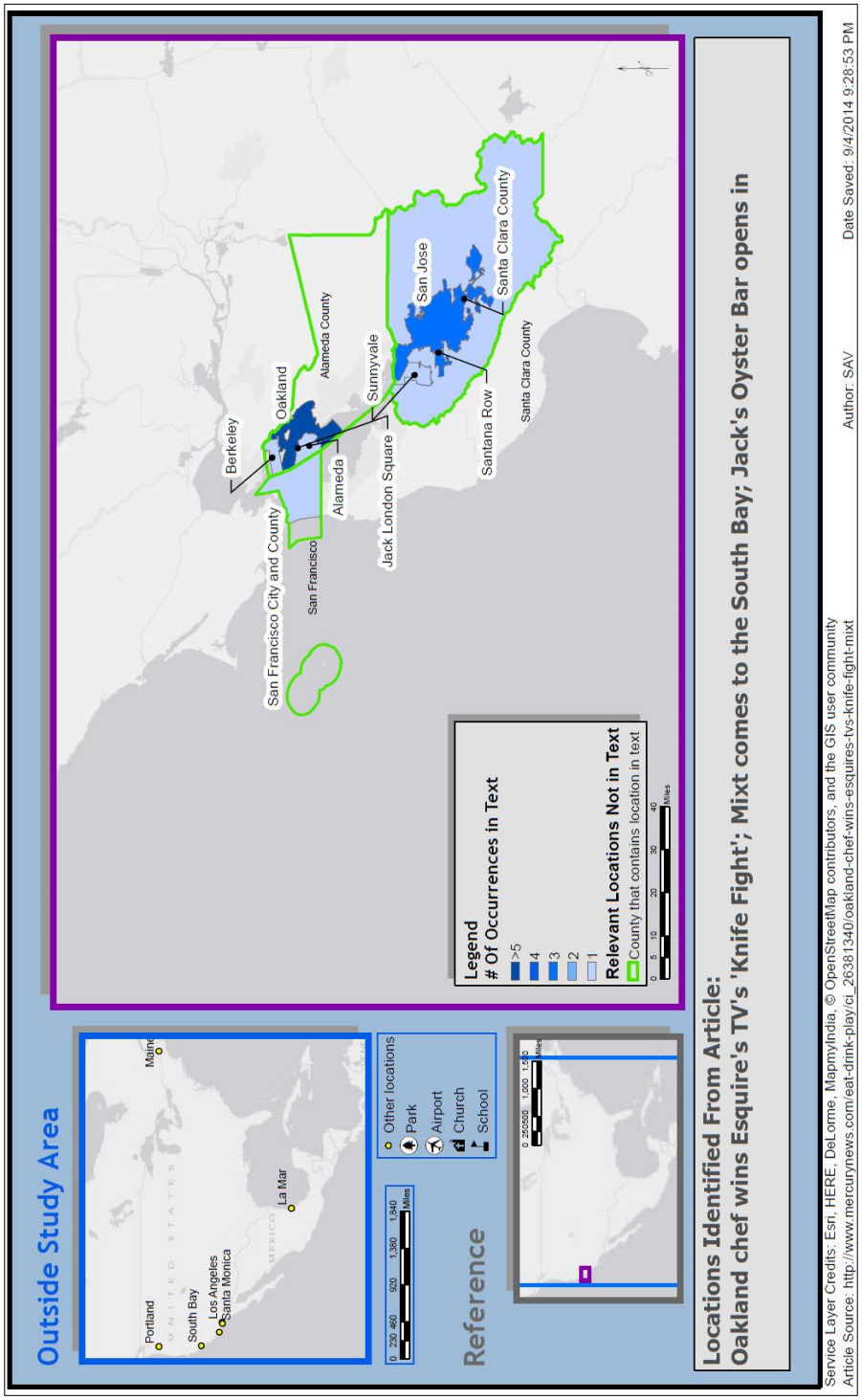


Figure A-14 SAV Generated Map with Manually Corrected PIG Results for Food Article

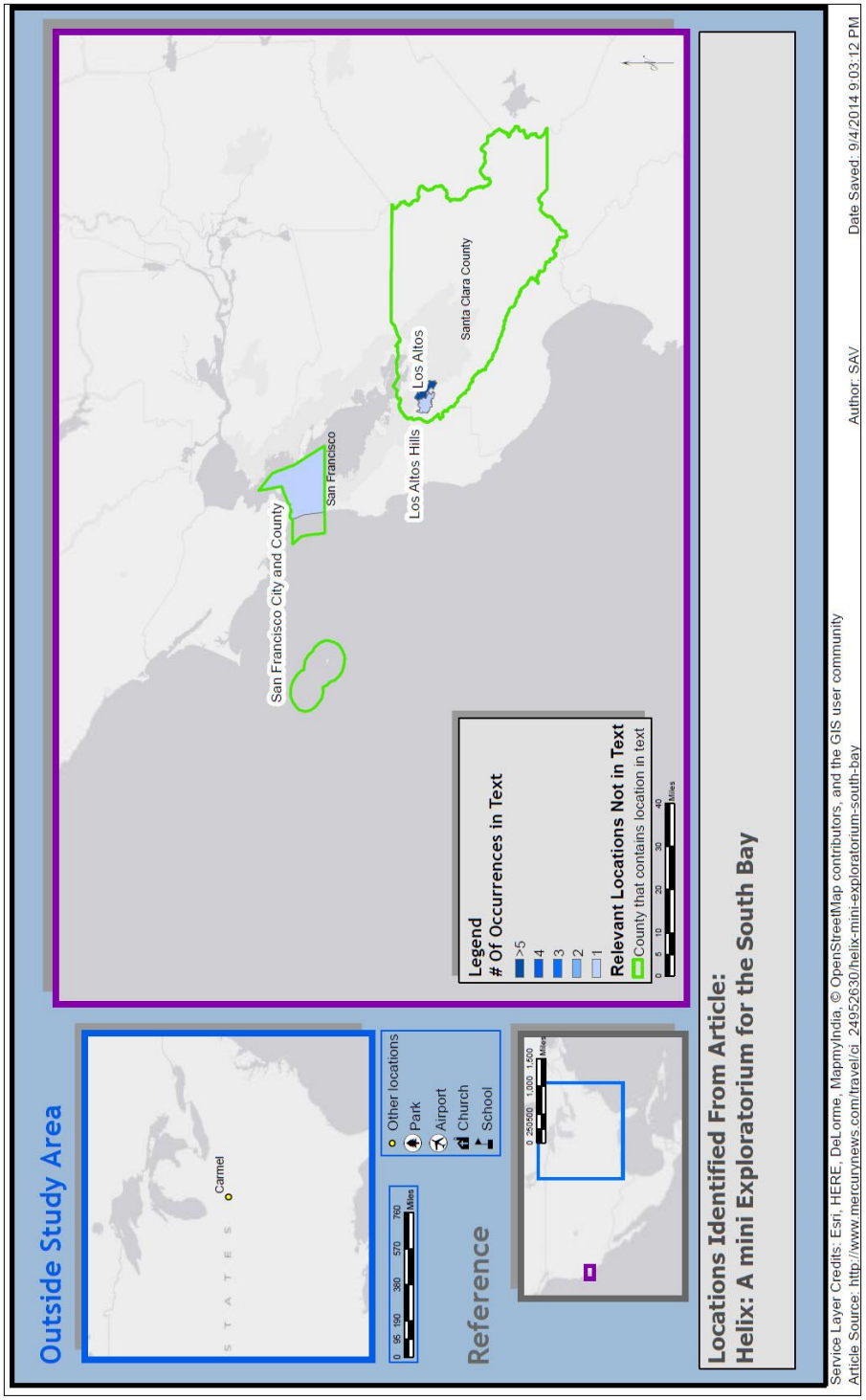


Figure A-15 Manually SAV Results Helix Article Map

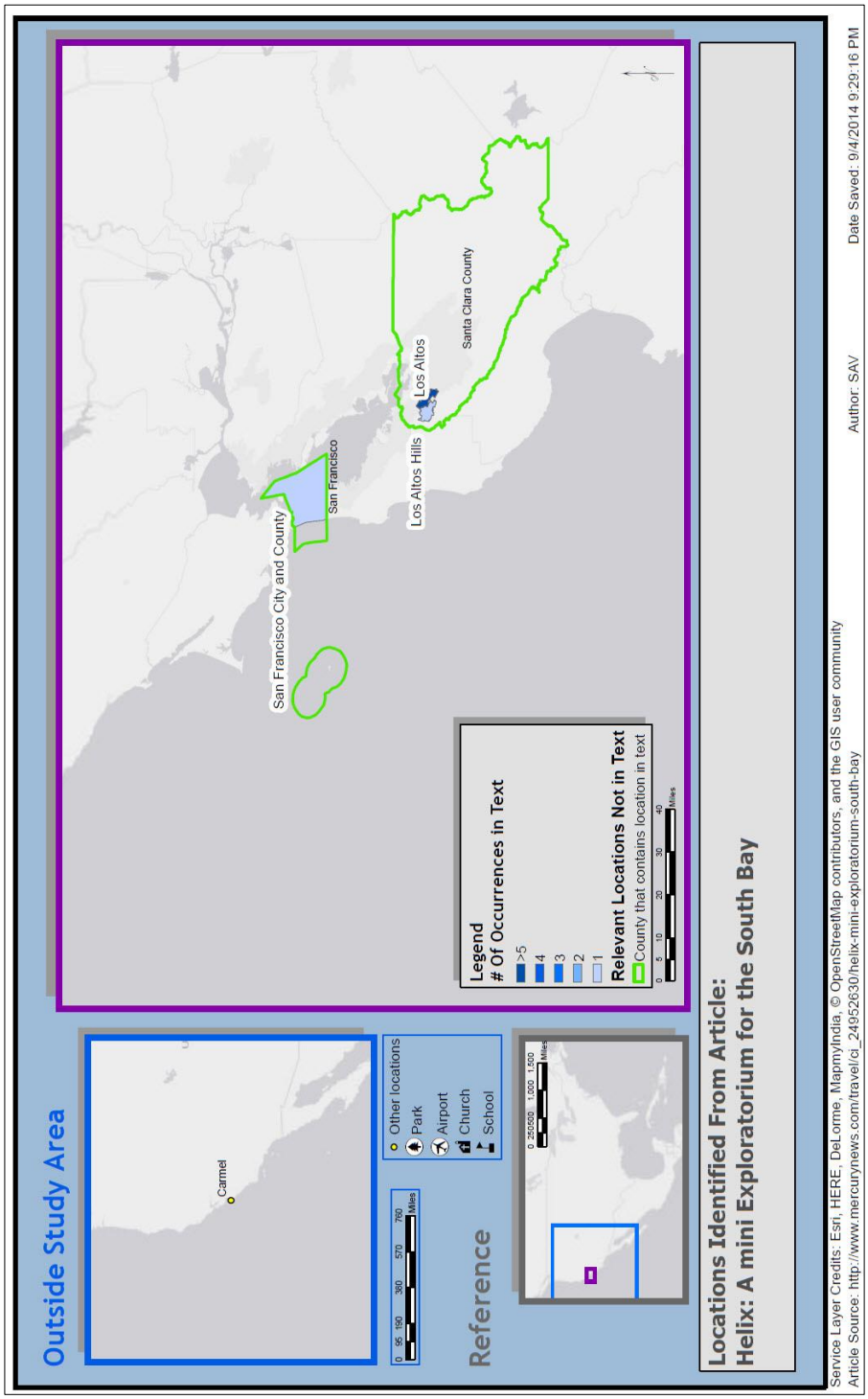


Figure A-16 Manually SAV Results Helix Article Map

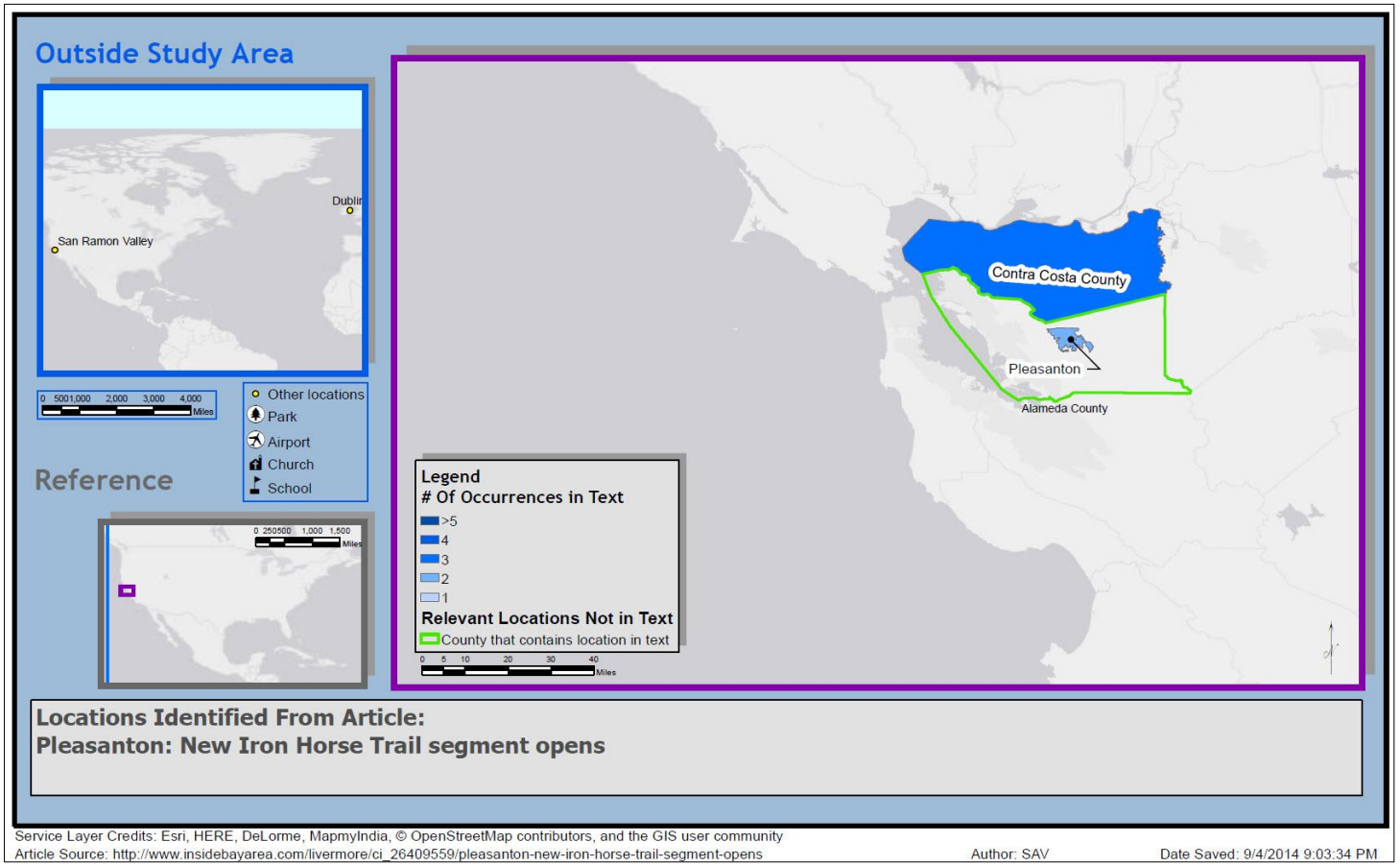


Figure A-17 SAV Results Iron Horse Trail Article Map

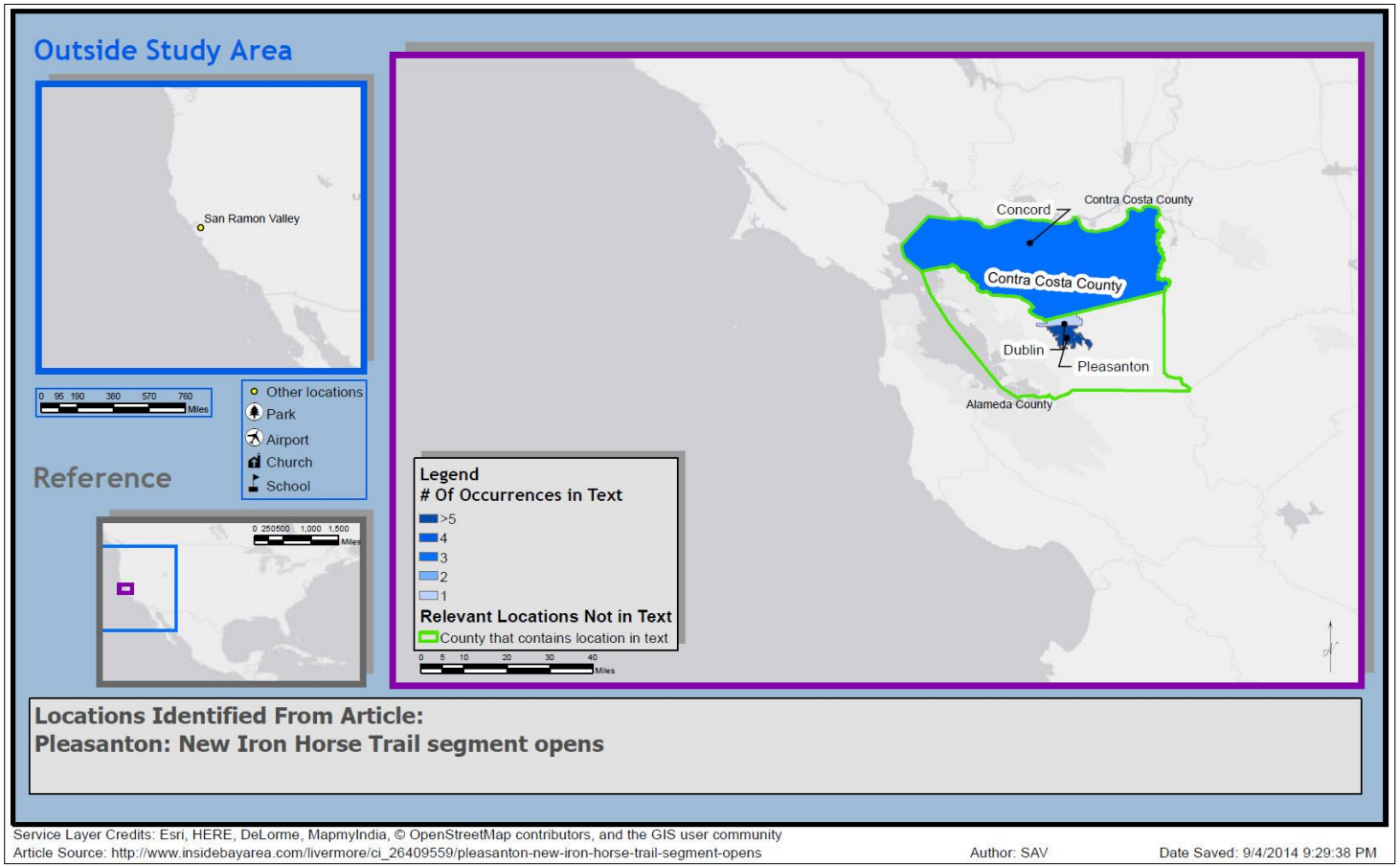


Figure A-18 Manually Corrected CLAVIN Results Iron Horse Trail Article Map

APPENDIX B: EXPERIMENT ARTICLE TEXT

Hate Crime Charge in Palo Alto Beating Case

A Redwood City man is accused of yelling racial slurs as he beat two others with a cane

By Robert Handa and Wire Services

Wednesday, Aug 27, 2014 • Updated at 7:02 PM PDT

A Redwood City man was arrested in Palo Alto last weekend for allegedly beating two men with a cane and yelling racial epithets at one of them, police said Wednesday.

Shane Patrick Collins, 26, was booked into the Santa Clara County Main Jail on suspicion of assault with a deadly weapon, felony robbery, misdemeanor resisting arrest and hate crime enhancements, Palo Alto police Lt. Zach Perron said.

At 11:58 p.m. Saturday, a police officer driving down the 400 block of Emerson Street heard a commotion from the back corner of a city parking lot, according to Perron. When the officer went to investigate, he noticed a man in his 60s lying on the ground in a fetal position with a suspect, who had blood on his hands and head, standing over the man and kicking him in the upper chest and head, police said. After the officer ordered the suspect to stop, the suspect fled on foot to an alleyway but the officer caught him and placed him under arrest, Perron reported.

An investigation revealed that the man in his 60s had been talking to a second man, who is in his 50s, when they were approached by the suspect, who was unknown to them, according to police. The suspect, who is white, then started repeatedly yelling racial slurs at the second man, who is black, Perron said.

“It was very evident to us that not only was this an 'assault with a deadly weapon,' but that that there should be a ‘hate crime' enhancement to it,” Perron said.

The victims told the suspect to go away and leave them alone, but the suspect then proceeded to assault the second man, who tried to defend himself by swinging his cane at him, police said. The second man then fell to ground and the suspect grabbed the cane, got on top of him and hit him at least once on the head with it, according to police.

The first victim, who is white, attempted to flee down the alleyway but the suspect chased him and struck him in the back of the head with the cane. When the man tried to defend himself, the suspect struck him again, breaking the cane over the man's forehead, police said. After the man fell down, the suspect began punching and kicking him until the officer arrived, according to Perron.

"The suspect was standing over him, holding on to a stairwell to keep his balance while he was kicking the victim in his head and upper chest,” Perron said.

One man suffered two cuts to his head and the other suffered one cut to his head, while the suspect also had a cut to his forehead, police said. The two victims were treated by medical personnel at the scene and released while the suspect was transported to a hospital to be medically cleared before being placed into jail, Perron said.

Rev. Jethro Moore, the president of the Silicon Valley chapter of NAACP said, if the allegations are true, Collins should be prosecuted for a hate crime. "If they hear about it and there is not strong, stern action, others will repeat it,” Moore said. “That's why it's important for us to handle it immediately after it happens."

The Palo Alto Police Department is wrapping up its investigation and it will be up to the district attorney's office to decide on the final charges.

Figure B-1 Hate Crime Article Text
Source: NBC Bay Area (Handa 2014)

San Jose: 16-year-old charged as adult in July homicide

By Mark Gomez

mgomez@mercurynews.com

Posted: 08/23/2014 06:50:53 AM PDT

SAN JOSE --A 16-year-old San Jose boy will be tried as an adult for his alleged role in the gang-motivated shooting death of a 22-year-old man in late July, according to the San Jose Police Department.

Carlos Enrique Vargas, 16, is the second person arrested and charged with murder and a gang enhancement in the July 28 slaying of San Jose resident David Escalera in the Gramercy neighborhood in East San Jose

He is the second teen to be charged as an adult in a homicide case this summer. Earlier this month, Marvin Garcia was still 14 when he was charged as adult in the stabbing death of an Oakland man in an alleged gang attack.

Vargas was in Santa Clara County juvenile hall on unrelated charges when police linked him to the Escalera case. His arrest was reported by police Friday.

He is the second suspect arrested in the Gramercy shooting; on July 29, police arrested 19-year-old German Alexis Arjona, who was also charged with murder and a gang enhancement.

According to police, Escalera was walking with his girlfriend on Gramercy Place in the middle of the afternoon when a group of about a half-dozen males confronted them. When he tried to avoid them, he was shot multiple times, police said.

A third suspect is still being sought: Homicide detectives also obtained a felony warrant for 19-year-old San Jose resident Humberto Bravo, who they believe was part of the hostile group. A photo of Bravo was not available, and police gave only a general suspect description that listed him as 5 feet, 9 inches tall, 130 pounds with brown hair and brown eyes.

Contact Mark Gomez at 408-920-5869. Follow him at Twitter.com/markmgomez.

Figure B-2 Homicide Article Text
Source: Mercury News (Gomez 2014)

San Ramon Man Arrested After Allegedly Firing into Ex-Girlfriend's House

Craig O'Sullivan, 62, of San Ramon was arrested on Saturday night

Posted by Autumn Johnson (Editor) , March 17, 2014 at 11:05 AM

By Bay City News—

A man was arrested after allegedly firing shots into his ex-girlfriend's San Leandro home on Saturday night, police said.

Officers arrested 62-year-old San Ramon resident Craig O'Sullivan shortly after they responded at about 10:45 p.m. Saturday to a report of a shooting in the 1200 block of Pacific Avenue near Thrasher Park. Police arrived at the residence, located about two blocks from the San Leandro BART station, and determined that the victim's ex-boyfriend had come to her home and discharged a firearm.

O'Sullivan fled the scene prior to police arriving but officers caught up with him as he was driving on Marina Boulevard. He was arrested and booked into county jail in connection with the shooting, police said.

Figure B-3 Shooting Article text
Source: Patch.com (Johnson 2014c)

51 Confirmed Measles Cases in California Prompts Officials to Issue Warning

Officials advise adults and children who will be traveling internationally to get the measles vaccine

Posted by Autumn Johnson (Editor) , April 08, 2014 at 05:53 AM

Health officials are advising people planning international travel to take precautions against measles due to a high incidence of the disease in California this year.

As of Friday, there had been 51 confirmed measles cases reported in California so far in 2014.

There had only been four reported cases by the same time last year, according to the California Department of Public Health. Four of the reported cases were in San Mateo County, four in Contra Costa County, two in Alameda County and one in Santa Clara County.

The rest of the cases occurred in Southern California. Most of the California measles cases have been contracted by people who were exposed to the disease while traveling internationally, including to the Philippines, India and Vietnam, or who came into contact with international visitors, according to the Monterey County Health Department.

The Centers for Disease Control and Prevention issued a travel notice for the Philippines in March due to more than 15,000 suspected cases of measles in that country between Jan. 1 and Feb. 15 of this year, including 23 deaths.

The CDC advises adults and children over 12 months of age who plan to visit the Philippines to get two doses of the measles vaccine 28 days apart for optimal protection. Infants between the ages of six and 11 months should get one dose of the measles vaccine before travel, according to the CDC. However, they will still need to get two doses of the vaccine when they are older.

Two doses of the measles vaccine provides near 100 percent protection from measles, according to the CDC. International travelers can check the specific CDC recommendations for their destination by visiting www.cdc.gov/travel.

Complications of measles include pneumonia, permanent hearing loss and death, according to the CDC.

http://sanramon.patch.com/groups/politics-and-elections/p/51-confirmed-measles-cases-in-california-prompts-officials-to-issue-warning

Figure B-4 Measles Article Text
Source: Patch.com (Johnson 2014a)

Family of UC Berkeley Student With Measles Contracts the Disease

Officials say they do not believe anyone else has been exposed

Posted by Autumn Johnson (Editor) , March 01, 2014 at 08:12 AM

By Bay City News—

Two relatives of a University of California at Berkeley student who contracted measles earlier this month have also caught the disease, Contra Costa County health officials said.

Neither the student nor the two family members were vaccinated against the measles, according to Contra Costa Health Services officials. Both relatives, who are Contra Costa County men in their 20s and 30s, voluntarily quarantined themselves in their homes after their relative was diagnosed with measles.

County health officials said it does not appear that anyone else has been exposed to the disease because of these cases.

However, Contra Costa Health Services' Communicable Disease programs chief Erika Jenssen said anyone who is not immunized is "very likely to get measles if they are exposed to the virus."

"This really underscores the importance of everyone getting vaccinated," Jenssen said. County health officials believe the UC Berkeley student contracted measles while on a recent trip to the Philippines.

BART users were put on alert about their potential exposure to the disease after the student used the transit system between the Downtown Berkeley and El Cerrito Del Norte stations during the first week of February.

The dangerous, highly contagious virus can be spread when an infected person coughs or sneezes, county health officials said. Anyone who rode BART between Feb. 4 and Feb. 7 is encouraged to look out for potential symptoms of measles through this weekend.

Measles symptoms can surface one to three weeks after being exposed and include coughing, runny nose, high fever and red, watery eyes. Two or three days after the fever begins, a rash develops on the face and spreads to the rest of the body, usually lasting about five to six days, according to county health officials.

A person infected with measles is also contagious for several days before and after the rash appears. Reported cases of measles both in the Bay Area and statewide have risen in recent months, according to the California Department of Public Health.

Source: http://berkeley.patch.com/groups/around-town/p/family-of-uc-berkeley-student-with-measles-contracts-the-disease

Figure B-5 Measles Berkeley Article Text
Source: Patch.com (Johnson 2014b)

Marin has second-highest rate of whooping cough in state, report says

By Janis Mara, The Marin Independent Journal,

Posted: 01/27/2014 07:57:57 AM PST1 Comment | Updated: 3 months ago

California saw a sharp increase in cases of whooping cough in 2013, and Marin had the second-highest rate of the highly contagious respiratory disease, according to a new report released by the state.

Nearly twice as many cases of pertussis were reported in California in 2013, a total of 1,904 statewide compared with 1,023 in 2012, the California Department of Public Health reported. The disease was once thought to have been all but eradicated.

With 173 cases, a rate of nearly 68 cases per 100,000 people, Marin ranks No. 2 statewide; Nevada County had the highest rate. The disease causes violent coughing, with coughing spells that can last as long as 10 weeks.

As to the reason for the increase, "it's unpredictable. It varies from year to year. We don't always know what determines the extent of a given outbreak, but we do know what we can do to prevent it," said Dr. Matt Willis, Marin County's public health officer.

"Vaccination can help prevent the spread of pertussis," Willis said. "We do have a lot of parents in Marin County who are hesitant about vaccines and we do know that we can protect ourselves by making sure every child is vaccinated."

Willis said the majority of 2013 Marin pertussis cases occurred in the months of May and June.

"We experienced a school-based outbreak that started in May and progressed until school got out in early June," Willis said. "In school, children are gathered in classroom settings and can transmit infection to each other.

The outbreak was not confined to any one area, Willis said.

"Most schools in Marin were affected. By the end of the year, 26 of our schools reported cases," Willis said.

The doctor said 160 of Marin's 173 cases were children younger than 19.

"Pertussis is spread easily in schools, it's highly contagious and it's a disease that affects children more obviously than adults. Adults get pertussis, but it's a less severe form," Willis said. "Infants are affected the most seriously with pertussis. In the 2010 outbreak of pertussis in California, 10 California infants died."

The 10 infants were the only pertussis-related deaths that year, according to Health Department information. None of the infants were in Marin, though the county had the state's highest rate of pertussis that year.

No one was reported to have died of the disease in California in 2011, 2012 or 2013.

The reason infants are more likely to die is that they are too young to have had the entire five-shot regimen recommended by the American Academy of Pediatrics. Also, their immune systems are still developing, the doctor said.

"We had a mini-outbreak from April through June of last year that we saw in our office," said Sara Koenig, a certified pediatric nurse practitioner with Tamalpais Pediatrics in Greenbrae. Tamalpais Pediatrics, which also has facilities in Novato, has 8,000 child patients.

"The majority of the kids we saw were teenagers, freshmen and sophomores in high school," Koenig said.

"It's important that everybody get vaccinated," Koenig said. While a vaccine can't guarantee that a child won't get pertussis, "if they do, the symptoms will be milder," the nurse practitioner said.

Figure B-6 Whooping Cough Article Text
Source: Sacramento Bee (Craft 2014)

Oakland chef wins Esquire's TV's 'Knife Fight'; Mixt comes to the South Bay; Jack's Oyster Bar opens in Oakland

By Jessica Yadegaran and Linda Zavoral

Mercury News

Posted: 08/25/2014 03:00:00 PM PDT0 Comments | Updated: about 5 hours ago

Jack's opens in Oakland: The long-awaited Jack's Oyster Bar & Fish House, has opened in Jack London Square to shellfish-happy crowds. Owners Rick Hackett and Meredith Melville (both of Bocanova) and Bocanova executive chef Peter Villegas (Campton Place, La Mar) have combined their talents to bring Oaklanders everything from salads and sandwiches to entrees and a raw bar in a chic, waterfront-inspired space.

The crowd at Jack's Oyster Bar & Fish House is proof that Oaklanders are hungry for a new seafood joint. (Nicole Kilian)Menu highlights include clam chowder with bacon; Maine lobster roll on a housemade pretzel bun; and squid ink carbonara with a slow-cooked egg and pecorino. The 100-seat restaurant, designed by architect Michael Guthrie and interior designer Ann Rockwell, features handcrafted chairs by Berkeley-based Wooden Duck furniture, with outdoor tables made from reclaimed wine casks and an oceanic mural by local artists Becky Carter and Michael de La Torre. Open from 11 a.m. to 10 p.m. daily. Details: 336 Water St., Oakland. http://jacksoakland.com.

Tossed to order: After serving more than 2 million pounds of local, organic lettuce in San Francisco and Los Angeles, the owners of Mixt Greens have decided to branch out. It's been a long time coming. Nine years ago, chef Andrew Swallow teamed up with his sister and brother-in-law, Leslie and David Silverglide, to create an ultra-green restaurant, meaning an eco-friendly eatery that specializes in made-to-order "farm to salad" cuisine.

Love hearts of palm? Green papaya? Lentils? Roasted beets? You can add any and all to your salad, then have it tossed with one of 14 gluten-free housemade dressings. If you don't need more decisions in your workday, there are imaginative Mixt combos and seasonal salads. Their first South Bay location opens Aug. 27, at Valley Fair in Santa Clara, and then they'll open next at Santana Row in San Jose. Although Swallow's other locations cater to lunchtime customers only, these Mixt Greens will stay open through dinner. Details: www.mixtgreens.com.

Hopscotch hero: In last week's Esquire Network's "Knife Fight," chef Kyle Itani of Oakland's Hopscotch was declared the winner over chef Jason Paluska (The Lark, Santa Barbara) in a live halibut challenge. Itani will celebrate his victory with Hopscotch fans through September by offering a poached Petrale sole (it's more delicate, he says) with miso sofrito collard greens to the menu, a dish inspired by two he presented to the judges to claim his "Knife Fight" victory.

Itani and Paluska were given three secret ingredients -- flowering cilantro, collard greens and live halibut -- and one hour to complete at least two dishes featuring the ingredients. Itani was declared the winner by judges Michelin-starred chef Josiah Citrin (Melisse, Santa Monica) and James Beard winner chef Naomi Pomeroy (Beast, Portland), along with host Ilan Hall. Hopscotch is located at 1915 San Pablo Ave, Oakland. http://hopscotchoakland.com.

Best 'cue: The inaugural Oaktown Throwdown team barbecue competition at Oakland's Art + Soul was a big hit, so expect a second annual, organizers say. The winners: Ric's Righteous Ribs, of San Jose, Grand Champion; Too Ashamed to Name, of San Jose, Reserve Champion; Bad S. BBQ, of Sunnyvale, People's Choice; Captain Kev's Competition Cookers, of Oakland, Backyard Barbecue; and Butcher's Daughter BBQ, of Alameda, Best Dessert.

Figure B-7 Food Article Text
Source: Mercury News (Yadegaran and Zavoral 2014)

Helix: A mini Exploratorium for the South Bay

By Tony Hicks

thicks@bayareanewsgroup.com

Posted: 01/20/2014 03:00:00 PM PST0 Comments Updated: 7 months ago

Mia Schmitt, 9, of Los Altos Hills, plays with magnetized washers at Helix in Los Altos, Calif., on Sunday, Dec. 15, 2013. Helix is a new community science

Amesha Banjara, 8, tests a batch of slime she made at Helix in Los Altos, Calif., on Sunday, Dec. 15, 2013. Helix is a new community science center

Letting an 11-year-old scientifically prove she has faster reaction times than I do by trying to catch a falling stick with numbers on it wasn't what I had in mind when I took my kids to the Helix Community Science Center in downtown Los Altos.

I got five rematches. I lost them all. My kid learned how to measure human reaction time with a fairly elementary piece of equipment she could probably construct herself. I learned I'm getting old.

Open since December, the center is the equivalent of a college extension program miles away from the big campus -- in this case, the Exploratorium in San Francisco. The center is in downtown Los Altos, which has some aspects of a smaller, less-expensive Carmel. It's fun, with the same type of adventurous, interactive exhibits as the Exploratorium, but the Helix center won't take much more than 90 minutes to get through, unless your kids start digging into the books and cool experiments they can take home (and you can pay for) in the extensive retail area.

In this case, size doesn't necessarily matter, especially since patrons are allowed to pay what they wish. There are 5,000 square feet to cover, and as usual, kids absolutely love the interactive scientific attractions, gizmos and gadgets. The Exploratorium and, by extension, the Helix center excel at having kids learn, while not necessarily knowing they're learning, which is why some parents -- like us -- like it so much. It's like putting salsa on vegetables. Make them taste good, and kids won't know they're doing something that's good for them.

Once my daughter took great delight in making me look bad, we moved on to the Chaotic Pendulum, which demonstrates how a simple set of three swinging pendulums creates a crazy amount of motion. My kids had no idea of the effect of wind on the ocean -- and maybe their parents didn't either -- until they spent some time with the Confused Sea, which demonstrates how moving air generates waves.

Then my 5-year-old found the sound exhibit, Gaussian Melody, in which three balls falling through patterns of pins generate random melodies. By the time she started writing songs to her melodies and singing for strangers, it was time for Dad to distract her with some other shiny science fun.

Emmy Weltsch, 2, of Los Altos, looks at how wind forms sand dunes with her father, Jerry Weltsch, at Helix in Los Altos, Calif., on Sunday, Dec. 15, 2013.

Watch the Depth Spinner long enough, then look at a wall and suddenly it looks like it's surging at you.Then there's the electromagnet ring toss, which allows kids (and me, for about 20 minutes) to build things using flat dime-sized magnets.

In addition to the 25 semi-permanent exhibits, there are monthly ones that focus on particular areas. January's was about light, which fascinated my kids no end. When you're 5 and someone gives you a fluorescent light and tells you to start drawing on a wall in a semi-darkened room, that's like proof that dreams come true.

The retail section was a lot more fun than I expected. Among the goodies was a camera that pinpoints which of your pores are sweating, therefore acting as a sort-of lie detector. I got away from that one fast, before my wife started asking questions. Kids also can take home crystal-growing sets, robotic toys, science experiments, erector sets, kits to build carnivorous insects and models exploring human anatomy. And lots and lots of books.

The Helix Community Science Center is operated by a one-year grant from Passerelle Investments, a Los Altos investment company. Add lunch and some shopping in the surrounding downtown, and it makes for a great centerpiece for a family day out.

Contact Tony Hicks at thicks @bayareanewsgroup.com.

Figure B-8 Helix Article Text
Source: Mercury News (Hicks 2014)

Pleasanton: New Iron Horse Trail segment opens

By Denis Cuff dcuff@bayareanewsgroup.com

Posted: 08/26/2014 12:09:32 PM PDT0 Comments Updated: about 23 hours ago

PLEASANTON -- Finishing the last Pleasanton segment of the 32-mile Iron Horse Regional Trail required overcoming many challenges -- millions of dollars to raise, lukewarm early political support and a creek and many streets to cross.

There were even unauthorized tennis and basketball courts a homeowners group built years ago on the county-owned path that was to become the trail.

It all came together last week, though, as the East Bay Regional Park District dedicated the $5.2 million, 1.6-mile trail segment stretching from the Dublin-Pleasanton BART station to Santa Rita Road. The off-road paved trail now extends 32 miles north to Highway 4 in Concord in Contra Costa County.

The paved trail provides a Pleasanton route to hike and ride directly to the BART station and through the Hacienda Business Park, making the city more connected to public transit and other regional trails in the system, officials said.

"This closes the last major gap between Pleasanton and Contra Costa County through an urban trail system," said Jim Townsend, the regional park district's trail development manager. "It is a significant milestone. Part of the big picture is that we're taking federal dollars not to widen a freeway, but to improve connections for hiking and riding."

The largest share of the trail costs -- $3.7 million -- was paid by a federal "Tiger II" grant awarded to projects that improve livability, environmental sustainability, safety and economic competitiveness. Extra rating points also were given to projects that foster partnerships.

The park district was among a small group of agencies that won such grants through a competition involving thousands of agencies. Of the other funding for the trail, $1 million came from bonds approved by East Bay Regional Park District voters; $350,000 came from Caltrans environmental funds, and $180,000 came from federal transportation funds.

To build the trail, park district contractors had to slice through the unauthorized basketball and tennis courts built along the county-owned route. In a negotiated deal with a homeowners group that built the courts, the regional park district agreed to pay $200,000 to rebuild the sports court in the Owens Valley Park in Pleasanton, park officials said.

"We want to be good neighbors," Townsend said.

Getting the sports courts built in a city park was a good deal for Pleasanton, said Susan Andrade-Wax, the city's community services department director. The city also benefits from the improved transportation options for people on foot and bicycles to travel in Pleasanton -- especially to the BART station, she added.

Pleasanton and Dublin officials were reluctant to support the Iron Horse Trail project during a big housing and business growth era in the 1980s, when trail construction began in the San Ramon Valley in Contra Costa County, recalled Bob Doyle, the regional park general manager.

But as the cities grew, the city leaders over time became eager supporters of the project, he added. Cooperation of the city and county was essential to bringing out the latest phase of the Iron Horse Trail, Doyle said.

Figure B-9 IH Trail Article Text
Source: InsideBayArea.com (Cuff 2014)