FREQUENCY AND PROXIMITY CLUSTERING ANALYSES FOR GEOREFERENCING TOPONYMS AND POINTS-OF-INTEREST NAMES FROM A TRAVEL JOURNAL

by

Scott D. McDermott

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

_____	Dr. Matthew T. Rice, Dissertation Director
_____	Dr. Nigel Waters, Committee Member
_____	Dr. Chaowei Yang, Committee Member
_____	Dr. Douglas Wulf, Committee Member
_____	Dr. Anthony Stefanidis, Department Chairperson
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs and Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science
Date: 4/28/17	Spring Semester 2017 George Mason University Fairfax, VA

FREQUENCY AND PROXIMITY CLUSTERING ANALYSES FOR GEOREFERENCING TOPONYMS AND POINTS-OF-INTEREST NAMES FROM A TRAVEL JOURNAL

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Scott D. McDermott
Master of Urban and Regional Planning
Florida Atlantic University, 2004

Director: Matthew T. Rice, Professor
Department of Earth Systems and Geoinformation Sciences

Spring Semester 2017
George Mason University
Fairfax, VA

DEDICATION

For my parents, who taught me that it is never too late to achieve one’s dreams.

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my adviser, Dr. Matthew T. Rice for his continuous support to my Ph.D. dissertation and research, and for his time and effort to guide me through the Ph.D process. Next, I would like to thank my committee members: Dr. Nigel Waters, Dr. Chaowei Yang, and Dr. Douglas Wulf for their constructive critique of my dissertation which broaden my appreciation to the different perspectives related to my dissertation. Finally, I would like to thank Ron Mahabir for the opportunity to discuss and critique my dissertation, and to Edna McClung for editing and proofreading my paper to improve the quality of my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1. Esri: Environmental Systems Research Institute
2. GIR: Geographic Information Retrieval
3. GIS: Geographic Information Systems
4. GISci: Geographic Information Sciences
5. NER: Named Entity Recognition
6. NLP: Natural Language Processing
7. POI: Points-of-Interest
8. VGI: Volunteered Geographic Information

ABSTRACT

FREQUENCY AND PROXIMITY CLUSTERING ANALYSES FOR GEOREFERENCING TOPONYMS AND POINTS-OF-INTEREST NAMES FROM A TRAVEL JOURNAL

Scott D. McDermott, Ph.D.

George Mason University, 2017

Dissertation Director: Dr. Matthew T. Rice

Travel journals are past accounts specific to the author's experience and are written to convey a geographically situated story. The story achieves its purpose by using points-of-interest (POI) names and toponyms considered significant by the author. Toponyms and POI names are placed by spatial, temporal, and human agent properties that define the geographic instances. The geographic instance gives an automation a framework to narrow the geographic focus and assign geographic attributes to a placename. Travel journals contain the geographic instances based on the experiences made at a specific time and location along a traversed route of a trip, which makes travel journals an ideal data source with which to conduct geographic information retrieval (GIR) studies.

This dissertation is a GIR study to georeference toponyms and POI names from a travel journal. Toponyms and POI names are georeferenced to locate where the author visited or what the author observed along a travel path. Toponyms and POI names are

subjected to semantic ambiguities, which can incorrectly place or omit a toponyms and POI names due to shared names by other geographic or non-geographic contents. GIR relies on algorithms to maximize the georeferencing of spatially sensitive data while minimizing issues related to semantic ambiguities.

Frequency analysis and proximity clustering minimize semantic ambiguities and georeference the toponyms and POI names to their correct locations, and are used to identify the toponyms and POI names along the route of the travel path. Frequency analysis identifies the primary and adjacent state names for each chapter of the travel journal that acts as a container for the subsequent toponyms and POI names. Proximity clustering groups the toponyms and POI names based on the distance to the cluster group's centroid. A cluster group with a significant number of toponyms and POI names contains the placenames that are more relevant to the travel journal. The use of frequency and proximity clustering analyses helps to narrow the focus of the geographic scope to select states and identify the toponyms and POI names that exist along the travel path.

The reliability measurements for this dissertation yield a precision rate of 88 percent and a recall rate of 30 percent. The precision rate is comparable to similar peer-reviewed studies and shows that the methodology described in this dissertation can assist in the GIR process. Issues exist due to name matching errors between the travel journal, geoparser, and gazetteers; temporal disassociations between the time the journal was written and the time this dissertation was conducted; omissions of POI names from the gazetteer; and incorrect tagging by the geoparser. Future studies are needed to provide

better name matching between the travel journal, geoparser, and gazetteers and on managing POI names to become integral to the GIR process.

EXECUTIVE SUMMARY

This study presents and demonstrates a combined approach using frequency analysis and proximity clustering to georeference the toponyms and points-of-interest (POI) names from a travel journal. Travel journals are historic accounts specific to the author's experience and are written to convey and describe all names and toponyms considered significant. The toponyms and POI names contained within the travel journal provide the data required to conduct a geographic information retrieval (GIR) study to georeference the locations recorded by the author. GIR relies on algorithms to maximize the success of georeferencing spatially sensitive data while minimizing issues that cause invalid results.

Written works contain ambiguities that cause the GIR algorithms to incorrectly identify placenames or ignore the placenames. Difficulties in georeferencing placenames occur because toponyms and POI names, by themselves, frequently provide little or no context to proper geographic placements. A conceptual framework is required to provide context using components that are efficient in narrowing the geographic focus to identify the proper location of a placename. The framework includes spatial, temporal, and human agent components. Knowledge of the geographic history (place of birth, residencies, employments, families) of the human agent, primarily the author, raises the relative probabilities of possible originations of the travel journal and places of interest to the author. Temporal knowledge of the trips made in the travel journal determines the proper gazetteer resources required for the named entity recognition (NER) process. Spatial data are toponyms and POI names included in the travel journals and gazetteers.

The conceptual framework provides focus, relationship, and an efficient means to georeference placenames. This study demonstrates the importance of formulating a conceptual framework to provide context for each toponym and POI name.

Toponyms and POI names are identified and categorized by a natural language processing (NLP) application and georeferenced by an NER process. The Stanford natural language processing toolkit is a linguistic parser and recognizer which is publicly available. The NLP application is tasked to identifying and tagging text as a location within the travel journal. The NER process matches and provides geographic attributes to the toponyms and POI names identified by the NLP application using a gazetteer and web locator. The result is a list of all possible combinations of georeferenced toponyms and POI names that are associated with one or more U.S. state names matched by a gazetteer. The final deliverable provided by the NLP and NER applications is a list of texts tagged as a location and associated with a geographic coordinate. The list of location-based texts is ready to be processed by the frequency analysis and proximity clustering to identify the actual spatial representation of the text.

This dissertation treats the travel journal as a "bag of words" ignoring grammar and word order and focuses on frequency analysis and proximity clustering to georeference a name. Frequency analysis and proximity clustering are used to provide context to the geographic text collected to minimize ambiguities and improve the placement efficiencies of toponyms and POI names. The frequency analysis counts the number of United States names that fully or partially contain the toponyms and POI names. The number of state names increases as more local toponyms (cities, towns) and

POI names are identified and matched. Frequency analysis identifies the state names related to the section of the travel journal reviewed that provides a geographic scope to narrow the name matching of placenames from the travel journals with placenames to a state. Clustering analysis alleviates issues where geographic confusion exists due to ambiguities, flashbacks, and personal comparison. Clustering analysis builds upon Tobler's First Law of Geography which states that "everything is related to everything else, but near things are more related to distant things." Clustering analysis groups all toponyms and POI names identified by the frequency analysis based on proximity. A primary cluster group is identified as the group containing a significant number of toponyms and POI names. More than one cluster group can be used to identify the toponyms and POI names if the state names in the primary cluster group exist in the subsequent cluster groups. Proximity clustering is applicable when the mode type for a travel journal is highly linear and the toponyms and POI names listed are identified along the route. The advantage of employing a combined approach of frequency and clustering analyses is that it provides a naïve approach to group toponyms and POI names. The disadvantage of this combined approach is that the toponyms and POI names must sufficiently exist within both the travel journal and gazetteer to provide correct data matches to georeference the placenames.

This dissertation experienced four significant issues that impacted the ability to georeference POI names. The first issue is the lack of POI names included in the gazetteers and web locators causing a failure of POI names to be identified. The second issue is the frequency analysis identifying one primary state for each section analyzed,

and from that one state all adjacent states were obtained. This presupposition that each chapter contains one primary state was too focused and was rectified by including states explicitly mentioned in the travel journal. The third issue, is the temporal disassociations of POI names and toponyms explicitly stated within the travel journal that was published in 1989, but written in 1987, to match the data contained within the gazetteer resources that were published in 2016. POI names are spatiotemporally dependent and analyzing a travel journal written in 1987 using a gazetteer or web locator in 2016, a difference of 29-years, will increase the false negative counts due to POI names being temporally sensitive and no longer existing in the current physical setting or gazetteer. Finally, the naming conventions of toponyms contained within the travel journal and gazetteer reduced the number of toponyms to be matched. The travel journal has cities and counties named "Alexandria" or "Arlington," but within the gazetteer those same geographic entities are named "City of Alexandria" or "County of Arlington." A cross-reference guide is needed to map the toponyms found in the travel journal with those in the gazetteer and a check within the dissertation's programming script to include prefixes such as "County of" or "City of" can assist in the gazetteer match process. When spatial and temporal components are missing from the conceptual framework, the number of false negatives in this study increased.

The dissertation's methodology incorporates six GIR components: 1) Preprocessing, 2) Geoparsing, 3) Candidate Placenames, 4) Georeferencing, 5) Precision and Recall, and 6) Visualization. Each component of the GIR process is essential to develop a methodology to identify, collect, georeference, and map the toponyms and POI

names contained within a travel journal. Preprocessing prepares the travel journal for geoparsing and georeferencing tasks; geoparsing identifies and groups all geographic related text; georeferencing adds spatial references to the toponyms and POI names collected; frequency analysis identifies the primary state name for a given chapter of the travel journal; spatial proximity clustering analysis groups all relevant toponyms and POI names; statistical measurements include precision and recall to measure the reliability of the toponyms and POI names collected; and data deliverables provide the visuals and graphs showcasing the analyses and end results. The methodology is a linear process to identify and place geographically referenced text from a travel journal onto a map.

This methodology returns a favorable precision rate and a recall rate that could be improved with additional work. The recall rate measures the application success in georeferencing all toponyms and POI names that are contained within the travel journal. The number of false negatives related to the inability to match most POI names, temporal disassociations, and naming conventions resulted in a 37 percent recall rate. The precision rate measures the application success in georeferencing all toponyms and POI names that are collected and stored in the bag of words. Precision rates asked of all the placenames that were georeferenced and given a place on a map, how many of those placenames were true positives? This study yielded a precision rate of 92 percent which is comparable to the precision rates from other peer-reviewed GIR studies.

Most GIR studies focus primarily on toponyms, and as a result, their precision and recall rates are based mainly on toponyms. This study has shown that POI names, when included in the analysis but listed as false negatives, impact the recall rate. Future

studies should provide more focus on georeferencing POI names. Managing and updating the temporal attributes of a POI name, e.g. date built, start and end date, assists in gazetteers management of POI names. What prompts a person to record geographically sensitive data into a travel journal, and what is the frequency of recording to one's journal? Can travel journals improve artificial intelligence to pre-determined travel plans and destinations based on previous trips made during a specific time. Future studies of these questions will depend on POI names and their inclusion into GIR studies. A major step for including POI names will be the development and management of a temporally sensitive gazetteer specific to POI names.

INTRODUCTION

Current technology allows individuals to transmit and record events observed or experienced during daily activity onto volunteered geographic information (VGI) platforms. A VGI platform acts as a conduit allowing individuals to share information, media, and observations related to an event's geographic instance designated at a specific temporal and geographic location. A geographic instance is a spatiotemporal snapshot that is of importance or interest to an individual, society, or a specific group defined by cultural or lifestyle backgrounds. This dissertation is interested in retrieving spatiotemporal places related to a geographic instance to promote the geographic information retrieval (GIR) process. GIR has a broad set of application domains and identifies geographic content to provide insight to the time and location of human activities. GIR informs individuals with disabilities of obstacles that hinder movement, navigation, and access to public space. These obstacles include barriers, detours, or construction activities that can be validated, quality checked, and added to accessibility maps for disabled individuals (Rice et al. 2011, 2012a, 2012b, 2013a, 2013b, 2014, 2015, and 2016). The United States law enforcement and intelligence communities review locations to reconstruct the movements of individuals (Cave 2016). Disaster response following the devastating 2010 Haitian earthquake was heavily influenced by geographic events related to the questions asked by Zook (2010): "Who needs help? And where?" Knowing the cognitive, social, and economic stimulus that influence personal attitude toward a travel destination allows the manipulation to a geographic instance to encourage trips to a specific destination (Um and Crompton 1990). In the examples mentioned, the

use of geographic information related to an instance of an event identifies geographic components that help formulate a response to questions that aid in personal, security, emergency, and trip responses. This study builds upon the general GIR process to provide a conceptual framework to identify geographic locations based on frequency analysis and proximity clustering to categorize toponyms and points-of-interest (POI) names.

This study presents and demonstrates a combined approach using frequency analysis and proximity clustering to obtain the primary states encapsulating the toponyms and POI names from a travel journal. Once obtained, this study isolates the toponyms and POI names, based on proximity, that represent the locations either visited or observed by the author of the travel journal. This approach builds a geographic hierarchy structure comprised of the primary and adjacent states and the toponyms and POI names contained by the states. Travel journals are used for this study as they document significant facts recorded by an author about a trip over a fixed period (Golob and Meurs 1986). Significant facts, as discussed by Golob and Meurs, are any events the author felt was important to transcribe into their travel journal. Each event consists of a spatial and temporal context of a travel event, and the human agent component. Collectively, these three event parameters provide a spatial-temporal-human context for a toponym and POI occurrence. A travel journal offers a multitude of events (trips, descriptions of a visit, and observations of interest to the author) containing a collection of toponyms and POI names. This wealth of geographic sensitive data makes travel journals an ideal source to research geoparsing and georeferencing toponyms and POI names. Research

advancements in geoparsing and georeferencing travel journals have direct relevance in geointelligence and terrorism analysis (Cave 2016), as well as novel geographic information applications such as Esri's Story Maps, where text, images, and geographic data are combined to present a user's georeferenced story. Similarly, National Geographic Society has had a longstanding interest in georeferencing and mapping the vast archive of photographic and textual material appearing in their magazine, which frequently describes geographic expeditions (Carroll 2006).

The wealth of potential location-based data comes at a cost to system performance and recall rates. The formally and informally named features in the United States consist of millions of toponyms and a much higher number of POI names, all of which could potentially exist within a travel journal. Table 1 displays a sample of the estimated number of U.S. toponyms.

Table 1 United States toponym counts based on 2014 U.S. Census.

Toponym Types	Total Counts
States	50
District and Territories	6
Counties	3,142
Incorporated and minor civil divisions	752
Incorporated and Census Designated Places (Including Puerto Rico)[1]	29,550

The United States has over 29,000 incorporated and Census designated places (CDP). The incorporated and CDP include cities, boroughs, towns, villages and

[1] https://www.census.gov/geo/maps-data/data/gazetteer2014.html (Under "Places" drop down)

unincorporated entities such as communities and commercial structures. [23] Figure 1 displays the distribution of incorporated places based on population within the U.S. which shows a significant number of places under a population of 10,000 (number of places = 16,470) to the remaining incorporated places with a population between 10,000 and one million (total number of places = 3,035). The U.S. has a total of 33,500 toponyms with most of the toponyms belonging to smaller incorporated towns.

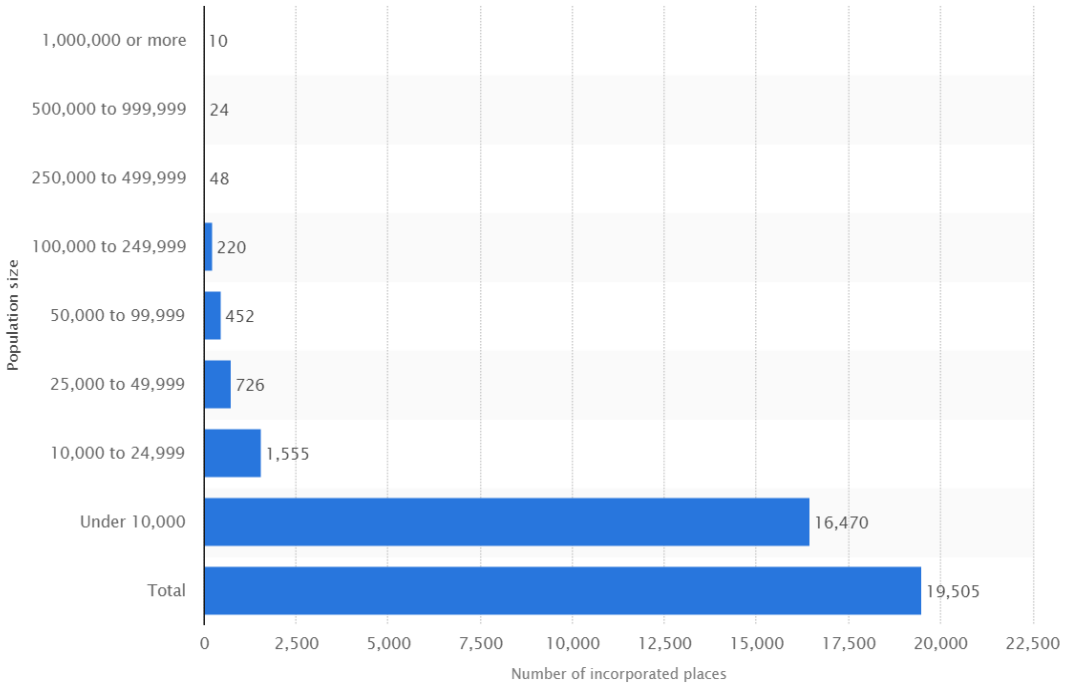

Figure 1 Number of cities, towns, and villages (incorporated places only) in the United States, 2015. Graph provided by Statista using U.S. Census Bureau data.

[2] https://www.census.gov/content/dam/Census/data/developers/understandingplace.pdf
[3] https://www.census.gov/geo/reference/gtc/gtc_place.html

Current GIR studies exist that extend their data sources to include local human developments (Gelernter and Balaji 2013). Table 2 displays the number of potential POI instances in the United States at more than 543 thousand POI destinations—businesses, restaurants, and commercial entities—and 131.7 million housing units. POI names are numerous and pose similar challenges experienced by toponyms to minimize ambiguities. Excluding housing due to its generic naming convention ("my house", "let's go home"), the POI developments pose a challenge to parse and reference the names in the GIR process due to the high numbers of potential names.

Table 2 Select sample of toponym counts based on U.S. Census 2012 Industry Economic Statistics[4].

POI Types	Total Counts
Housing	131.7 million
National Parks	398
Museums, historical sites, and similar institutions	7,125
Public Educational Places	105,000
Restaurants	431,000

Previous GIR studies have shown high precision using NLP and NER association for geographic text (see Table 3). The second column of Table 3 contains the reported statistical precision, recall, and F-score (the harmonic means of precision and recall that measures accuracy) from different peer-reviewed studies. Applying GIR techniques against toponyms and POI names for the United States is a challenging task. The United States contains close to 33,500 toponyms in addition to half a million POI developments that could serve as potential human destinations. The number of placenames provides a

[4] http://www.census.gov/econ/snapshots/index.php

significant challenge for GIR studies due to semantic geo/geo and geo/non-geo ambiguities. Geo/geo ambiguities occur when a name can represent multiple toponyms such as "Springfield, Massachusetts" or "Springfield, Illinois." Geo/non-geo ambiguities occur when a name can represent a toponym, person, or organization such as "Washington" which can be a U.S. toponym or a personal or pet name. Defining a context that can be implemented to a machine-level process can minimize ambiguity issues.

Table 3 Examples of statistical accuracies from previous GIR studies.

Application or Processes Name	Statistical Accuracy Outcome	Source
Nominator	92% (precision rate)	Wacholder, Ravin, and Choi 1997
Web-A-Where	80% - 91% (precision rate)	Amitay et al. 2004
Word frequencies	80% (accuracy)	Verma, Vieweg, Corvey, Palen Martin, Palmer, Schram, Anderson 2011
Supervised machine learning	73 - 85% (accuracy)	Hu, Ge
Gazetteer Classification	78.5% (predictive accuracy training	Garbin, Mani 2005
Supervised machine learning using GeoCLEF for Travel Blogs	86.7% (precision rate; recall: 38.1%)	Nanba, Ozaki, Ishino, Taguma, Kobayashi, and Takezawa 2009
Microtext Geoparser	.90 (F-Score) .99 (precision, toponyms) .94 (recall, toponyms)	Gelernter and Balaji, 2013

Because travel journals record a personal series of events that can occur in micro settings, POI names become extremely important to decrease the geographic footprint of an event contained within a toponym. Geoparsing and georeferencing POI names and toponyms from a travel journal allow an examination of personal travel experiences and a description of the human interactions that enrich those locations. Travel journals were not prominent in GIR studies reviewed for this dissertation, but travel journals provide toponyms and POI names that make travel journals advantageous for the GIR process. The accessibility of travel journals relies on the availability of electronic documents and the popularity of mobile devices. Mobile devices are becoming highly important for referencing a placename and are expected to be the primary means of network communicative connectivity by 2020 (Kaplan and Haenlein 2010). Travel journals can benefit from the popularity of mobile technology allowing people to become more geo-aware and communicative of their location-based experiences. The advancement of technology has led to the development of a social media and technology platform that encourages and showcase the sharing of geographically situated and geographically referenced data, pictures, text, and media.

Goodchild (2007) popularized the concept of volunteered geographic information (VGI) as the general populace acting as human sensors voluntarily providing geographic content made possible by the advancements in technologies. Meaning that people are afforded a platform to freely contribute geographic sensitive data for others to view and comment. Technological advancements provided the required platforms and network infrastructures to allow people to describe events at a more personable scale. Humans

acting as sensors, with the aid of portable devices, provide a means for a much smaller and more focused spatial reference defined by human developments that are geographically associated to the administrative and political boundary. Travel journals and blogs are a medium that stores geographic sensitive data that is specific to an individual experience. The geographic sensitive data stored within the travel journal offers a chance to capture a smaller geographic footprint contained by or intersected to a boundary of a toponym. The use of smaller geographic micro settings will provide an extension to the toponym hierarchy to improve the toponym resolution.

Travel journals display similar characteristics as VGI, but travel journals are not bound by technical platforms. Travel journals offer geographic sensitive data based on personal experience and can be in the form of a hard or electronic copy. A travel journal is a document that will contain at least one spatial attribute to define the location visited or observed (Amitay et al. 2004). The knowledge that a document will have at least one spatial attribute is important for the retrieval, disambiguation, ranking, and storing of administrative and political toponyms from textual narratives (Larson 1996). The collection of states, subordinate toponyms, and POI names creates a toponym hierarchy that is defined by frequency analysis and proximity clustering.

Identifying the primary state for each section of a travel journal, and the location of toponyms contained within a travel journal is possible by frequency analysis. Frequency analysis is common in GIS and has shown results for landslide (Pradhan and Youssef 2010, Cervone et al. 2016), vehicle trips (Crane and Crepeau 1997), and drought severity (Loukas and Vasiliades 2004). Frequency analysis collects and counts all

possible toponym candidates with the anticipation of identifying all state names related to the document. Frequency analyses are dependent upon the availability of the data source within a gazetteer to determine the primary state name for a section of the travel journal. Candidate toponyms are georeferenced to capture all potential states. Frequency analysis is highly dependent on the quality of the names found in the travel journal and the management of the gazetteer to match to those names. Should any toponyms be counted as a false negative or positive, it can disrupt the frequency analysis by lowering the count of a valid toponym.

Frequency analysis does have another fault as it does not account for humans associating a place based on past experiences (flashbacks) or comparing a current geographic location to a different geographic area. Flashbacks and the human desire to compare one location to another can interfere and disrupt the frequency analysis required to determine the primary state name. Places mentioned in the travel journal but not relevant to the actual location can increase the production of false positive data, or data noise, to the analysis. An additional method is needed to minimize data noise related to flashbacks or toponym comparisons.

Banu and Karthikeyan (2013) described clustering as a preferred method for most georeferencing analysis. Even in cases where noisy data is persistent the input data derived from clustering can minimize that noise. Clustering will not eliminate all data noise from the cluster group as there are some degrees of uncertainty as defined by Halkidi et al. (2001). What makes clustering of toponyms and POI names most reliable is the grouping of similar toponyms and POI names based on proximal distance from the

cluster group centroid into one cluster group. This proximal relationship fits Tobler's First Law of Geography (Waters 2017). The data noise and irrelevant locations will be placed into different cluster groups separating potentially irrelevant candidates and grouping the preferred candidates. The collection of toponyms closely related to the travel journals provides some degree of confidence that the data obtained are valid to the sources used.

This dissertation demonstrates the use of frequency and clustering analyses to provide a list of toponyms and POI names best describing the spatial attributes for a travel journal. Travel journals are records of events containing geographic sensitive datasets composed of spatial, temporal, and human agent properties. By developing a mechanism that will adhere to these three properties, the risk of ambiguities is minimized, especially when a high number of toponyms and POI names are found within the document. There are additional issues contained within the travel journal that include flashbacks and toponym comparison that can provide unwanted false positives. Although no formal means were developed to minimize toponyms related to flashbacks this study will attempt to use flashbacks to favor a more positive outcome when identifying the scope and location of a placename in the document. This study will display how the issues of flashbacks and toponym comparison are minimized by employing frequency and proximity clustering analyses to georeference toponyms and POI names from a travel journal. It is the intent of this study to formalize and implement a process that will geoparse and georeference a formal travel journal, minimize the data noise, and properly place the toponyms and POI names defined in the document.

To evaluate the travel journal using frequency analysis and proximity clustering there are four presuppositions. The first presupposition, is the use of the travel journal as "a bag or words." A bag of words ignores grammar and word ordering within the travel journal allowing a focus on georeferencing text resembling toponyms and POI names. The second presupposition is the location of the travel origination as a known location specific to the author. The origination is assumed to be a known location representing a place of residency, employment, or interest to the author. A collection of placenames that are relevant to the author (birth place, employments, and residencies) can identify the origination. If no match is made of the author's trip origination in the travel journal a manual review is required. The third presupposition is that all toponyms and POI names will be contained within the contiguous United States. The knowledge that toponyms and POI names are contained within the United States will provide focus for the analysis to restrict the locations within the extent of the US. Last, a presupposition is made that the travel is linear with a vehicle as the primary mode type. A linear travel route allows an assumption that the author will cross adjacent boundaries between two states and assumes the toponyms collected will be contained by one or more state names that share boundaries. The four presuppositions guide the analysis, which views the travel journal as a bag of words with toponyms contained within the United States and a linear travel pattern.

Hypothesis

The intent of this dissertation is to list the characteristics of a travel journal and display how frequency and clustering analyses can successfully georeference toponyms and POI names. Travel journals are used because they contain spatial, temporal, and human components that create a geographic instance contained within an event. The events portrayed by a travel journal are highly specific and act as identifiers to locations. To provide a sensible structure that defines the event contained within the journal, the three components are required to develop a hypothesis:

Toponyms and POI names contained within a travel journal are correctly identified and placed in their geographic location based on geographic instances defined by ***spatial****,* ***temporal****, and* ***human agent*** *properties.*

The dissertation's goal is to automate the process of georeferencing placenames which includes the travel journal starting point, collection of toponyms and POI names, identifying the states that contain the toponyms and POI names for a chapter, proximity clustering to group toponyms and POI names relevant to the chapter, and precision and recall. This automation greatly benefits complex tasks where large archives of text are run through the process. One example of a repository that can benefit from the automation is the National Geographic repository which contains 125 years of archive magazine materials. Magazine articles spanning 125 years are too complex and lengthy to process manually. Additionally, the intelligence community would benefit from the

same automation. Relevant surveillance and geointelligence analysis can be automated, and later manually reviewed where the author of such documents includes location-based text. The speed of the automated processing will provide immense benefit to geoparsing and georeferencing through many documents.

A geographic instance is a location containing an event and defined as the intersecting sets of properties composed of a spatial, temporal, and a human agent component. All three components from an event are required to effectively encapsulate and isolate a geographic toponym and POI name, and provide the proper context anchoring the geographic named entity. The hypothesis is an optimistic approach to develop a conceptual and practical framework that will identify and retrieve local POI names and toponyms from a travel journal, and to expand the description of the names to include geographic coordinates. The presuppositions of the hypothesis are based on literature reviews commonly associated with geographic textual retrieval and associations often found for toponyms.

Conceptual Framework

Travel journals are significant facts recorded by a respondent about a trip made over a fixed period (Golob and Meurs 1986). Personal accounts of a trip or series of trips made for varying purposes are maintained within a travel journal and often contain geographic components. The geographic components are described by its travel characteristics (Crane and Crepeau 1998), trip purposes (Guensler and Bachman 2001) and social interactions (Silvis et al. 2006). Travel characteristics, trip purposes, and social interactions are all events that can lead to geographic content. These geographic related events contain geographic sensitive data that can be utilized and placed onto a map.

All travel journals comprise three main components that can be categorized as spatial, temporal, and agent (Guensler and Bachman 2001; Crane and Crepeau 1998; Silvis et al. 2006; and Golob and Meurs 1986). Figure 2 shows space, time, and a human agent as the main contributors defining a geographic instance contained by a travel journal. The three main components of the travel journal (and geographic components of a travel journal, discussed above) reflect the three elements of a relational geographical framework (meaning, space, and social relations), as defined by Sack (1997) and summarized by Holt-Jensen (Figure 3, 2009). The interactions between meaning, space, and social relations are existential, linking the "genius loci" (spirit of a place) with "habitus" (social sense of a place) through physical nature. People relate to a geographic location and provide meaning to that location through personal experiences.

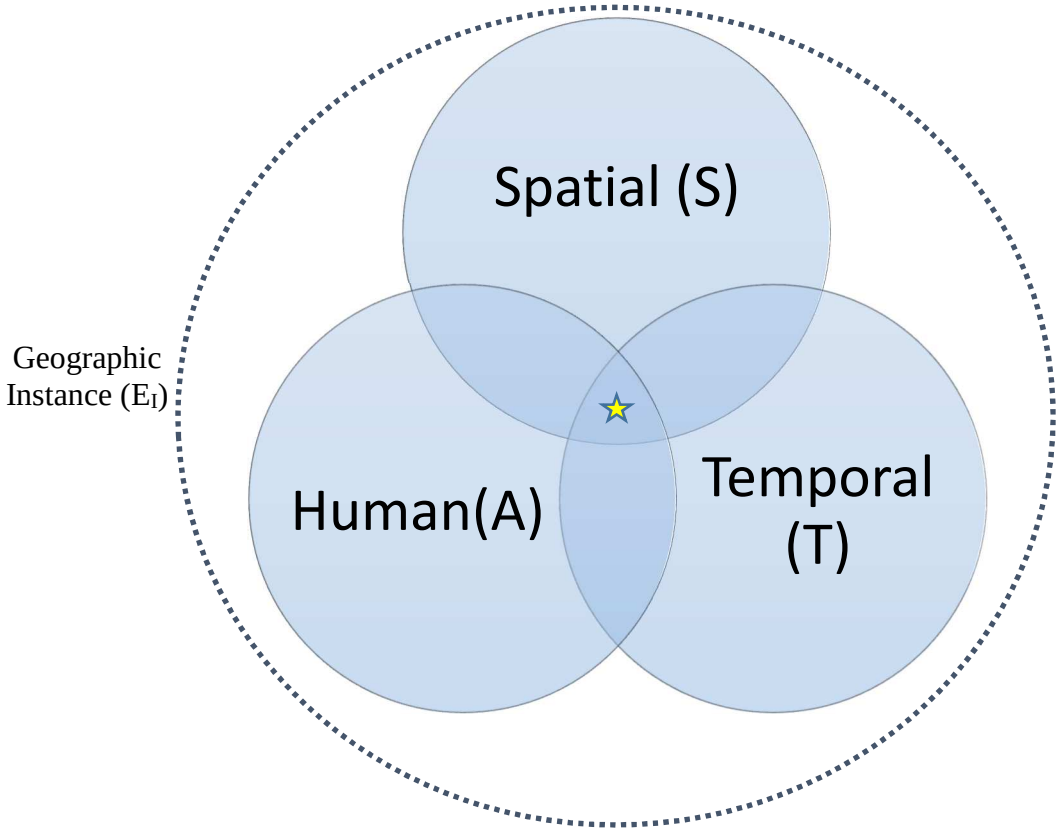

Figure 2 Event Requirements. The "star" represent the geographic instance (E_I). The three components displayed focus, a fully developed relationship, and efficiency for the study. $E_I \in S \cap T \cap A$

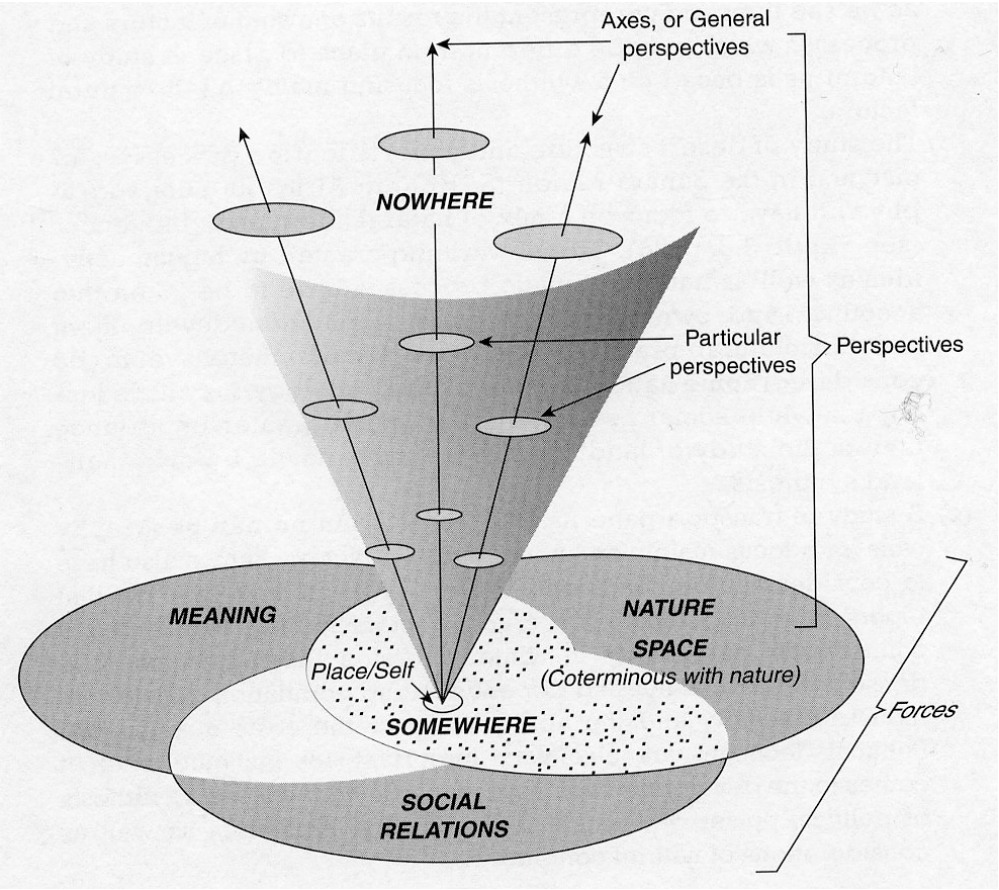

Figure 3. Relational Geographic Framework, from Sack (1997) and Holt-Jensen (2009)

Spatial Components

Studies exist discussing the use of toponyms in a GIR process, but not much focus has been applied to POI names. POI names represent mainly human developments that are geographically characterized and confined to a smaller scope than toponyms. There are varying descriptions of the ecological levels of human developments, but for this study a "micro system" best characterizes human developments as:

> …*Complex relations between the developing person and environment in an immediate setting containing that person (e.g. home, school, workplace, etc.)* (Bronfenbrenner 1977).

Bronfenbrenner's definition continued to define a setting as:

> …*a place with particular physical features in which the participants engage in particular activities in particular roles for particular period of time (*Bronfenbrenner 1977).

Both terms for micro systems and settings denote spatial and human agents. The setting are human developments, and human developments are physical places with specific characteristics for people to engage in their activities such as studies, eating, and living. The characteristics of a human development include home, school, and workplace, which are much smaller than political or administrative boundaries. The two characteristics define the types of development that can improve the toponym resolution. The setting of human development is extended to describe a POI name, a term borrowed and most commonly associated to a location-based tracking system (Zheng et al. 2009). For this dissertation, POI names are correlated with and connected to a specific human development that is significant to the author of the travel journal.

Human developments provide a smaller scope and geographic footprint to a document, and the lack of focus to include POI names within the GIR framework is disconcerting. Previous studies (Drymonas et al. 2013; Gelernter and Balaji 2013) demonstrate how POI names can contribute to the GIR process, but POI names are

susceptible to ambiguity issues as experienced by toponyms. Ambiguities and human phenomena such as flashbacks can disrupt the placement process of geographic names collected. Travel journals are often plagued by the concept of geographic indicators (near, adjacent to, part of), geo/geo or geo/non-geo usages, and geographic names. Linguistics phenomena, such as Lasersohn's (1999) "pragmatic halo," apply truthfulness to a location of a respondent when that respondent is not at that location, but is anticipated to be at that location within an appropriate time. When a POI name is disambiguated and correctly georeferenced the use of POI names are useful to narrow the geographic footprint for the travel journal.

Agent (Human) Components

User-contributions are a dominant source of geographic data and real-time events used by socio-technology media. Goodchild (2005, 2007) stated the advancement in network technology and the inclusion of social interactions popularized the concept of volunteered geographic information (VGI). The continuous evolution of web-based applications supports a real-time communication among its participants and encourages new and existing personal relationships (Bargh and McKenna 2004). The availability of a real-time communication medium and the likelihood that people have the motivation to contribute and maintain geographic sensitive data are driving the social construct attributed to geographic and temporal knowledge. User contributions and the motivation defining VGI can be extended to include geographic information exemplified in travel journals.

VGI is real-time and temporally sensitive (Goodchild and Glennon 2010; Kaplan and Haenlein 2011; Rice et al.2012). VGI is contributed with both desktop computers and mobile devices, which facilitate dynamic, interactive, georeferenced multimedia communication. Dynamic and interactive platforms allow groups of people with similar interests to post geographic sensitive data that are publicly available. Locations of traffic incidents (Medina et al. 2017), restaurant reviews, and disability obstacles, are displayed on a common platform available to everyone.

VGI actions occur online where a platform exists for people to genuinely reflect their thoughts (Yarkoni 2010) or provide their own self-narratives (Hirsh and Peterson 2009). Travel journals reflect personal thoughts and narratives and can be found both online or offline and have some common attributes with VGI. Travel journals account for specific personal experiences made at a time and place along the traversed route of a trip. The availability of geographic sensitive data with the author describing an event based on an actual visits or observations can contribute to inform the public of their historical or current geographical settings. The value of user contributions and their participatory role in aggregating information is an important concept related to VGI. Data from travel journals—when aggregated with crowdsourcing, personal, or publicly available platforms—can accommodate or answer geographically related questions. The emergence of VGI, and the related human components are reviewed comprehensively by Elwood et al. (2012).

Temporal Components

VGI shows strength in dynamic situations where explicit temporal items are of importance during emergency situations like forest fires (Kebler et al. 2009) or micro-blogging, which provides a real-time geospatial analysis of people's reactions and opinions, as manifest through social media (Stefanidis et al. 2013). Even historic journals or records of events can rely on explicit temporal elements (Grover et al. 2010). When time data is provided explicitly, an assumption is made linking the sites to the time identified; when no time is provided, an indirect or implicit means of generalizing the temporal analysis is required based on the local geographic entities identified and an assumption about linear, sequential travel.

Temporal elements are strongly encouraged to be integrated within gazetteers to represent changes (Hill 2006) and answer when an event took place (Gey et al. 2010). Temporal information is explicit, in the case of direct temporal expressions, or implicit, in the case of metadata and ordinal text data offering some time perspective to an event or existence of a place (Peregrino et al. 2012). Explicit temporal information defines the actual time and date stamps that can be retrieved within the pages of the document, forum comment headers, hashtags, or web URLs. Implicit temporal data are retrieved by metadata if a timestamp of that toponym exists within the gazetteer. Explicit temporal data are advantageous for social media, news articles, or micro-blogging websites where specific date and time indicate exactly when an event occurred or a narrative was written. Implicit temporal data are advantageous for chronological ordering where date and time are not used but contained within a metadata resource or described as an ordinal

narrative. Both explicit and implicit temporal events assist in developing a chronological ordering for a given spatiotemporal occurrence.

Travel Journal Data Source

The travel journal used in this study is "The Lost Continent: Travels in Small Town America" (1989) by Bill Bryson, and readily contains the spatial, temporal, and human agent components. Bryson's travel journal is a 28-chapter document that describes his encounters and observations during his cross country trip to different regions of the United States. This journal was selected for analysis because it contained the following:

- Singular human agent for most events.
- Linear travel behavior.
- Explicit use of toponyms and POI names.

Bryson's travel journal recounts his personal experiences while traveling across the United States. In the travel narrative, Bryson also discusses past personal life events as well as historical events and information about a location. Flashback and memory associations related to a specific site can potentially deviate from strict personal encounters with a location. Flashback is an instance of analepsis (Bae and Young 2008) which acts as a literary device to deviate the reader's comprehension process (Jakobson, R 1960). Flashback acts as a foregrounding mechinism to semantically shift or break the main story line (Mukarovsky 2014). The deviation caused by a flashback to move it to the foreground of the story, purposely disrupts the literary chronology to attract the reader's attention and provide a backstory to an event that occurred during Bryson's

travels. The actual events during Bryson's travel are chronological, but the intermittent flashbacks invoke a personal reason or perspective to a location.

Bryson's journey throughout the United States is linear, seperated into two parts, east and then west, and with the primary mode type of transportation being a vehicle. The mode type and linear travel allows an assumption that the author will travel on a roadway that crosses geographic space and administrative/political geographical units with specific associated toponyms, progressing to geographical units immediately adjacent to the geographical unit the author is currenly describing. The linear travel allows analysis to focus on a geographical structure with specific state names that are adjacent to one another rather than all state names. Restricting the number of toponyms not only limits the number of states for review but also improves the performance of the georeferencing process.

Finally, a major advantage of using Bryson's travel journal is that it offers explicit use of toponyms and POI names. Explicit names of toponyms and POI will minimize the issues related to vagueness or alias names, but there are instances where name matching between text and gazetteers do not yield any positive results. Gazetteers focus on toponyms defined by an authoritative source, e.g. the U.S. Board on Geographic Names, and it is difficult to locate a gazetteer that contains names not managed by an authoritative source, such as POI names. POIs are temporally sensitve and can be destroyed, go out of business, or undergo a name change which gazetteers frequently do not record. Hill (2009) suggests that gazetteers should incorporate time: "A gazetteer that does not have a temporal component cannot represent that changes that happen through

time to placenames, spatial footprints, relationships, and other characteristics of named geographic places" (ibid. 121). It is expected that not all POI names will be located due to the 27-year difference between the time the journal was published (1989) and the time this study took place (2016). Despite the time difference between the travel journal and this study, some POI names are expected to be georeferenced and will help narrow the focus of toponym resolution to a more local level.

Bryson's travel journal provides explicit toponym and POI names, follows a linear traverse behavior pattern using a vehicle mode type, and places focus on both himself as a traveler and the sights he experienced during his travels. Along his journey, Bryson continues to provide detailed accounts of his personal experiences. These experiences are often specific to a place along the route, and his decisions to visit these places are predetermined based on varying factors. The factors that determine Bill Bryson's route and places visited can be influenced by external or internal means. The inclusions of events based on geographic settings, temporal references, and the nationwide inclusion of toponyms offer a high availability of data for this study.

Methodology

Leidner and Lieberman (2011) outlined six components to process textually encoded spatial data. These components were modified slightly to conform to the dissertation implementing a frequency analysis and proximity clustering (see Figure 4): 1) Preprocessing, 2) Geoparsing, 3) Candidate Placenames, 4) Georeferencing, 5) Precision and Recall, and 6) Visualization. The process to georeference spatial data from

a textually encoded travel journal defines the means to georeference POI names and toponyms from the travel journal and place them onto a map. The program is written using Python 2.7 with the data stored in XML and JSON format. This will take advantage of Python's English-like programming structure to allow readers to understand the computational process.

In this study, the toponyms and POI names inherit the coordinates from the gazetteer and web locator resources. Once the coordinates are assigned, a spatial proximity clustering analysis groups related toponyms and POI names based on their distance to the cluster group's centroid. Next, statistical measurement based on precision, recall, and F-Score will measure the quality of the application. Longley et al. (2015) provide an excellent overview for the traditional usage of the terms accuracy and precision, within the geospatial domain. Recognizing some differences in the usage of the term precision, this study employs the term consistent with other GIR studies, where the precision is the ratio of all true positive data against the number of false positives collected. Precision may also be thought of as a measure of positive predictive value. Recall is the ratio of the true positive data against the number of false negatives. Precision denotes the success of placing toponyms with what was obtained, and recall denotes the number of potential candidates missed. The F-Score is alternatively known as the F_1-score or the F-measure and is a measure of the accuracy of binary classification (Gelernter and Balaji 2013). It is often described as the harmonic mean between precision and recall (Grover et al. 2010). Once all toponyms and POI names from the travel journal are analyzed through this, they are mapped and visualized for review. The

expectation of this study is to provide a formal approach to georeferencing POI names and toponyms within a travel journal.

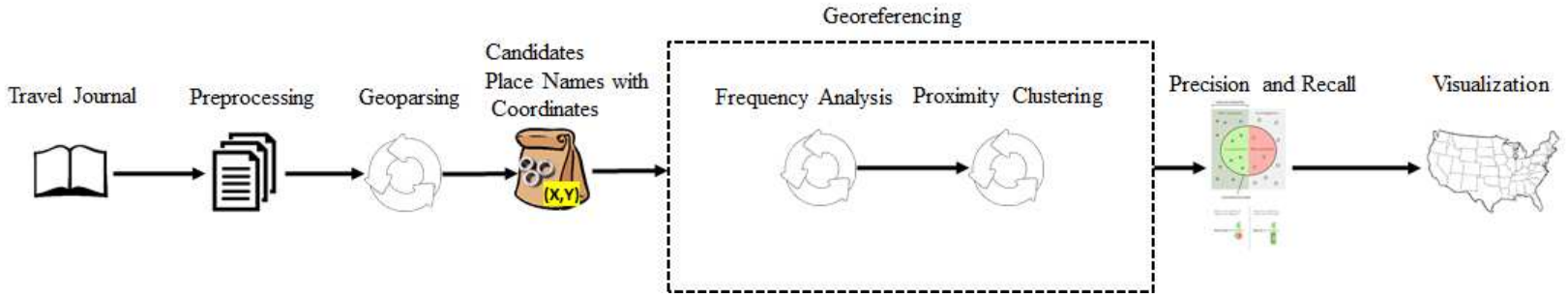

Figure 4 Outline of the general GIR process.

Results

A summary of the results is provided here as a guide and preview for the dissertation, following from the previous introductory summary of the methodology used to generate results. Inside the cover of Bryson's book is an artistic, cartographic rendering of the route traveled by the author (Figure 5). For comparison, Figure 6 is the computational outline provided by this dissertation plotting the POI names and toponyms for each chapter. The numbers in the figure represent a chapter of the travel journal. The points within the polygon represent the georeferenced toponyms. Chapter 17 (dashed polygon) displays an area missed from the analysis. Instead of displaying toponyms in New York, Pennsylvania, and Ohio, the analysis went from New York to Des Moines, Iowa (D1). This was a result of insufficient toponyms identified from the travel journal. The dash line (D2) displays the missing chapters not analyzed (Chapter 27 and 28) due to an error related to the Python name matching and georeferencing application. The two figures have strong visual correlation, with some noticeable discrepancies.

The study shows a favorable precision of 88 percent (see Table 4). The recall rate, 30 percent is significantly lower than expected due to the high number of false negatives related to inconsistent naming variations between the gazetteer and travel journal and limited POI names listed in the gazetteers.

Table 4 Precision and Recall results.

Precision, Recall, and F-Score	
Reliability Measurements	Results
Precision	0.88
Recall	0.30
F-Score	0.52

The recall was lower than anticipated due to the naming variation between what was used in the travel journal (e.g. “Alexandria”) and what was used in the gazetteer (e.g. “Town of Alexandria). A secondary cause of a low recall was the temporal disassociations between the time the travel journal was published (1989) and the time this study was conducted (2016). The current gazetteer lacks POI names in 1989. The results increased the number of false negatives and lowered the recall rate. Finally, POI names are limited in gazetteers making it difficult to georeference all POI names. Improving gazetteer management to include POI names would increase the recall rate for this study.

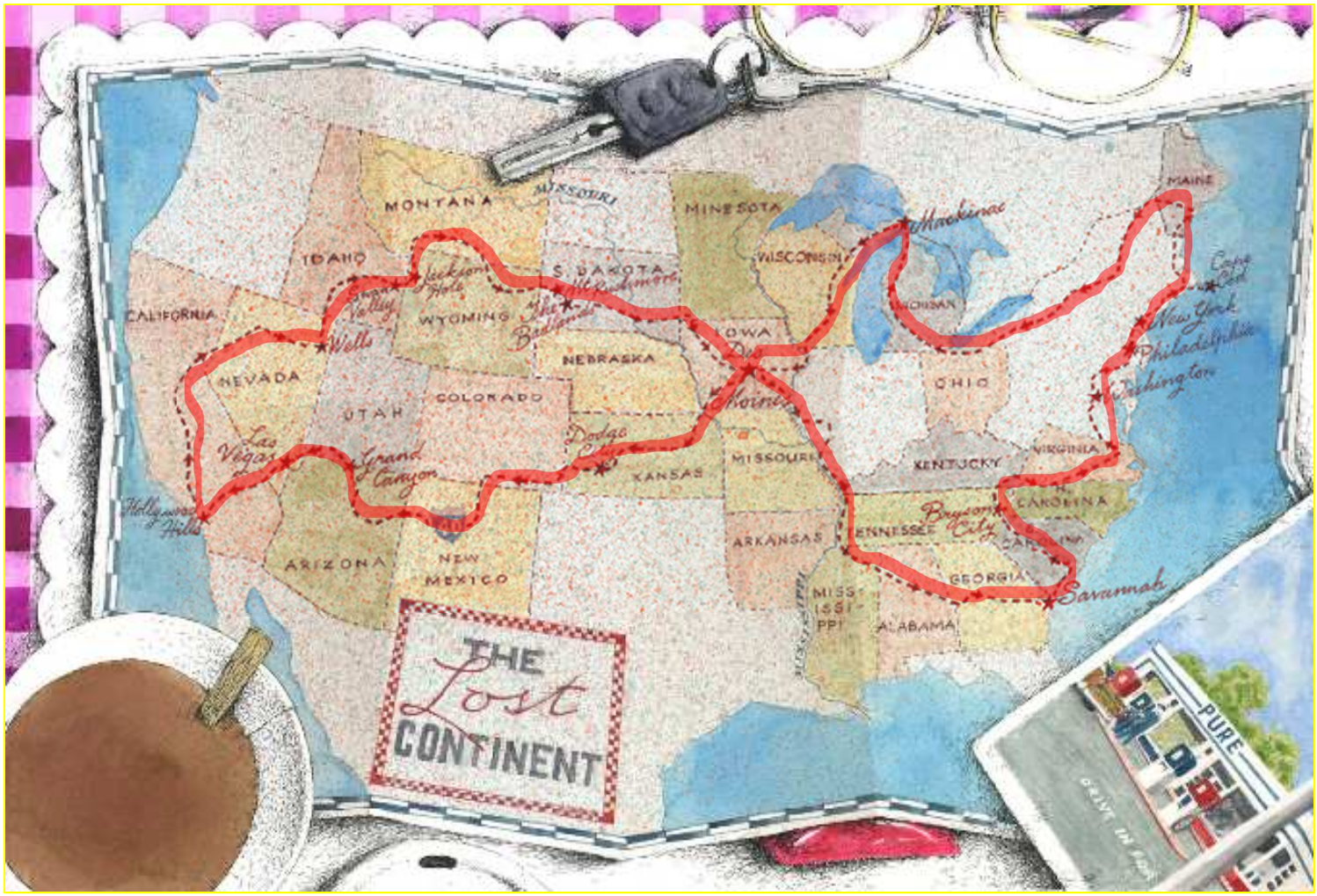

Figure 5 Outline of Bryson's travel across the United States. Image from "The Lost Continent: Travels in Small Town America" (Bill Bryson, 1989).

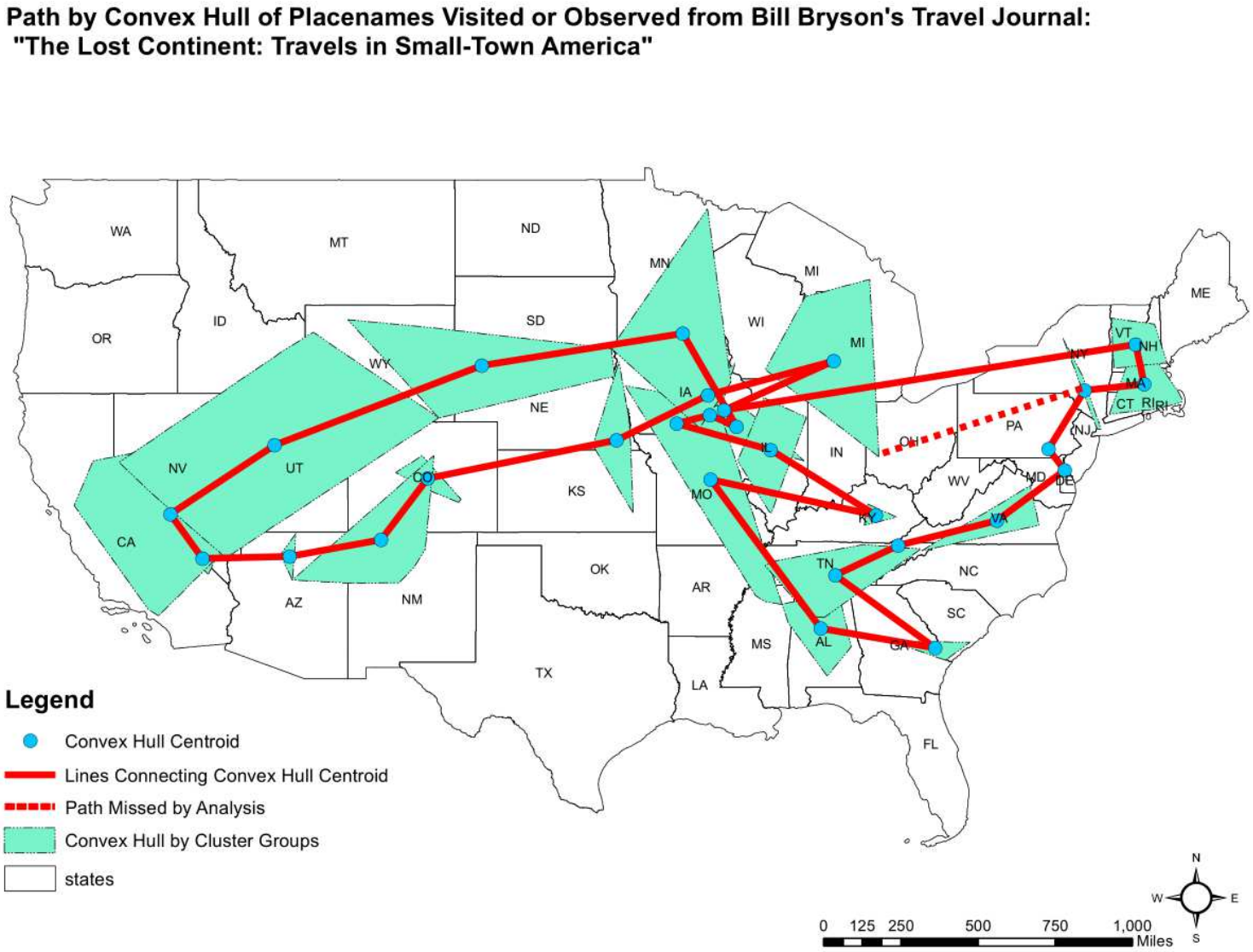

Figure 6 Final output outlining Bryson's trip. The dashed line depictit an error related to missed toponyms.

Another issue that impacted the georeferencing of placenames was the author's frequent use of flashbacks and the resulting spatiotemporal displacement of sections of the travel journal. Flashbacks and toponym comparisons constituted a major problem when validating the correctness of a georeferenced toponym. Chapters 16 and 17 in Bryson's "The Lost Continent" (The title is misleading as the travel journal is contained within a single country) are a good example of a flashback and spatiotemporal displacement, where the mapped travel trajectory shows a connection, via a dotted line, between New Hampshire to Iowa. The dotted line in Figure 6 was a result of the author's memory of his hometown baseball players from Iowa. In Chapter 17, the author is in Cooperstown, New York, discussing baseball and reminiscing how his home state (Iowa) contributed to the sport. The result of this memory occurrence related to a state not adjacent or contained in the geographical scope of the chapter causes the application to view Des Moines, Iowa, and not Cooperstown, New York, as the valid toponym for Chapter 17. The application resolved the travel pattern discrepancies in the next chapter by recognizing the explicit state names in the travel journal and matching the toponyms and POI names to those explicit state names. The study attempted to minimize these issues by employing frequency analysis and proximity clustering to georeference toponyms and POI names in travel journals. The dissertation's high precision shows that this study can correctly place toponyms and POI names and alleviate the impact caused by false positives.

Conclusion

Studies that exist in the GIR field examine the means to georeference toponyms from various electronic and non-electronic documents and images. This dissertation employs a combination of frequency and proximity clustering analyses to aid in defining a narrow geographic scope to minimize ambiguity issues that often plague GIR studies.

This research intends to focus on formulating a conceptual framework based on the definition and components used to define a travel journal. The conceptual framework validates the POI names and toponyms contained within a travel journal. The three components (spatial, temporal, and human agent) are required to confirm the identity of a toponym and POI name found within the text. A geographic instance of an event will provide the proper reasoning for the placenames by elaborating the existence of the toponym and POI names for a given time observed or visited by a human agent. Once a text is confirmed as a placename that name can be affixed to the end of a taxonomy tree.

This dissertation borrows heavily from previous studies that apply georeferencing techniques for political and administrative toponyms. How this study differs is in its emphasis on the importance of a conceptual encapsulation to provide context describing the spatial content. By emphasizing the importance of the conceptual framework, the focus shifts the importance from a singular word to the importance of a group of words to formulate an anchor for a placename. When the conceptual framework approach is successfully implemented the analysis leads to a result confirming the identity of a local geographic entity from a textual narrative.

This dissertation delivers on its intention to identify toponyms and POI names within a spatial, temporal and human agent context to georeference the name. The viability of toponym or local POI names for location analysis and temporal approximation is paramount when attempting to narrow the geographic footprint of an event. The success of this study is dependent on identifying a geographic instance, locating the components, and georeferencing the toponyms and POI names.

LITERATURE REVIEW

Geographic information retrieval (GIR) focuses on providing access to information that is spatially and geographically oriented for indexing and retrieval (Larson 1996). GIR references unstructured and structured geographic data from documents, providing context with the geographic data and identifying the location associated with the document's content. This study builds upon the GIR process with a focus on georeferencing geographic data from a travel journal through means of a frequency analysis and proximity clustering.

Travel journals are literature devoted to travels, tourism, social communications, or historic events and from another perspective, function as a container of toponyms and local Points-of-interest (POI) names. Travel journals contain significant facts recorded by a respondent (human agent) about a trip (spatial location) made over a fixed temporal range (Golob and Meurs, 1986). Travel journals record personal accounts or observations that occurred on a trip or series of trips made for varying purposes and provide travel characteristics (Crane and Crepeau, 1998), trip purpose (Guensler and Bachman) and social interactions (Silvis et al. 2006). These characteristics determine the geographic setting or behavior related by time, distance, mode types, number of trips, and inter-personal interaction.

The travel journal used in this study is Bill Bryson's *The Lost Continent: Travels in Small-Town America* (Bryson 1989) where Bryson traveled by car to places throughout the US. During his travel, Bryson described and named states, cities, towns, highways, parks, hotels, museums, monuments, and restaurants visited or seen along the way.

Toponyms and POI names are prominent throughout Bryson's journal, providing a mechanism to geo-reference the narrative. The challenge faced by many peer-reviewed studies are the semantic ambiguities resulting from vast numbers of toponyms and POI names that exist and can be duplicated, or interchangeable with other proper names. This literature review documents the challenges involved in the GIR process to georeference toponyms and POI names due to semantic ambiguities and identifying the true location of a placename.

Toponym and POI Names

Toponyms are placenames that are recognized by toponymic authorities (national, state, or local governments) and are the primary way to refer to a place (Kemp 2008). Toponyms are associated with city, region, and state features with a name that is recognized. Toponyms also include other features with names that are recognized such as schools, monuments, parks, and administrative government buildings. The authoritative placenames are used by gazetteers to provide a dictionary or indices of placenames (ibid). Gazetteers associate placenames to a geographic location and categorize them by feature types (country, states, cities, monuments, parks). Toponyms are authoritative and recognized by governing institutions, while POI names are placenames that are non-authoritative and less formal.

The challenge within the GIR process is to develop an algorithm to define a context for a POI name. Peer-reviewed studies, such as the Perseus Project (Smith and Crane 2001), Nominator (Wacholder et al. 1997), Web-A-Where (Amitay et al. 2004) or

SPIRIT (Silva et al. 2006), have detailed algorithms to identify and index authoritative toponyms, but little on POI names. Peer-reviewed studies with emphasis on POI names include Petar (Li and Sun 2014) which tag and inventory POI names, and studies relating POI names to a location based on geographic proximity or containment of a known toponym (Drymonas et al 2011; Gelernter and Mushegian 2011). The focus on POI names improves the toponym resolution to a larger scale.

POI names are less formal placenames that are not created or managed by an authoritative source. POI features are mainly commercial or private in nature that include hotels, restaurants, grocery stores, retails, and gas stations. Gazetteers are more focused on authoritative names, but databases do exist that contain POI names but mainly for commercial purposes such as advertisement and location based services. Toponyms for authoritative and POI placenames exist in different databases where gazetteers are more formal, recognized, and authoritative in nature, and databases containing POI names are more dynamic, and commercial in nature.

POI Inventory in the United States

Table 5 shows the number of toponyms and the land mass contained in the United States of America (US). The U.S. consists of nearly 2.0 percent of the world's total land and water masses, and 4.4 percent of the world's total population (2016 count from CIA World Factbook: 324 million)[5]. Within the United States are distinctive political

[5] https://www.cia.gov/library/publications/the-world-factbook/geos/us.html

boundaries consisting of 50 states and six districts and territories.[6] Contained within the states are 3,143 counties and equivalents (independent cities, boroughs, and parishes)[7] and 19,531 incorporated towns and cities[8] (of which 73 municipalities contain a population of 250,000 or more). The combination of U.S. counties, cities, and townships, states, and territories result in 22,730 political entities. The number of U.S. entities makes it susceptible to issues related to semantic ambiguities.

Table 5 United States basic inventory of land, population, and toponyms.

United States of America Inventory		
Inventory Item	Total	Percent of World
Land and Water Mass (million sq. km.)	9.8	2
Population (millions)	319	4.4
States	50	-
Territories	6	-
Counties	3,143	-
Cities and Townships	19,531	-

Placing a toponym in its correct location is challenging, due to the frequent presence of geo/geo and geo/non-geo ambiguities (Amitay et al. 2004; Cave 2016). Geo/geo ambiguities occur when a name can represent multiple toponyms such as "Springfield, Massachusetts" or "Springfield, Illinois." Geo/non-geo ambiguities occur when a name can represent a toponym, person, or organization. Table 6 displays an example of geo/non-geo names where county names can also represent historical U.S.

[6] http://www.census.gov/prod/cen2010/doc/dpsf.pdf
[7] http://www.census.gov/popest/data/counties/totals/2012/CO-EST2012-alldata.html
[8] https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk (Incorporate and CDP)

figures (Washington, Lincoln, and Jefferson). The name Washington is found in 30 states and can represent the first U.S. president, a state, a district, a monument, or other names. When the name Washington is used, a context is required to understand the text meaning, and if a toponym or POI name, its location.

Table 6 Eleven county names that are shared by other states. The "Count" field is the total number of the name that exists in the US.

County	Number of States where the Name Exists
Washington	30
Jefferson	25
Franklin	24
Jackson	23
Lincoln	23
Madison	19
Montgomery	18
Clay	18
Monroe	17
Marion	17
Union	17

How many POI names exist within the US? The U.S. Census Bureau provided a nationwide snapshot of industrial and commercial economic statistics[9] (see Table 7) by category types. These counts do not tell the entire story regarding the number of entities existing within the US, but they do show that millions of potential POI-related features exist, and conceivably, nearly all of them have a location and a name. The U.S. Census Bureau estimates that 7.2 million industrial and commercial facilities exist. These

[9] http://www.census.gov/econ/snapshots/index.php

establishments do not all represent a concrete physical location with a name, but the U.S. Census list does offer an insight into the large number of food, retail, businesses, and recreational services existing in the US. Equally disbursed, the number of establishments across the 3,143 counties is about 2,302 establishments per county. A significant number of potential POI names exist within the U.S. and its counties. The number of potential placenames increases both the complexity and magnitude of georeferencing the toponyms and placenames.

Table 7 Summary of POI entity counts within the United States. 2012 Data

Number of Institution Types and Employees Within the United States		
Institution Type	Counts	Total Employees
Mining, quarrying, and oil and gas extraction[10]	28,643	903,641
Utilities[11]	17,578	649,988
Construction[12]	729,345	7,316,240
Manufacturing[13]	296,605	11,268,906
Wholesale trade[14]	420,014	5,947,657
Retail trade[15]	1,062,646	14,705,820
Transportation and warehousing[16]	213,805	4,316,392
Information[17]	134,652	3,206,226
Finance and insurance[18]	470,081	6,056,417
Real estate and rental and leasing[19]	354,731	1,926,027
Professional, scientific, and technical services[20]	854,274	8,142,951
Management of companies and enterprises[21]	52,380	3,065,905
Administrative and support and waste management and remediation services[22]	385,314	10,217,859
Educational services[23]	68,276	653,409
Health care and social assistance[24]	830,813	18,587,467
Arts, entertainment, and recreation[25]	124,347	2,092,370
Accommodation and food services[26]	662,487	12,006,316
Other services (except public administration)[27]	528,371	3,456,130
TOTAL	7,234,362	114,519,721

[10] http://www.census.gov/econ/census/pdf/7121.pdf
[11] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=22
[12] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=23
[13] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=31-33
[14] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=42
[15] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=44-45
[16] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=48-49
[17] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=51
[18] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=52
[19] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=53
[20] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=54
[21] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=55
[22] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=56
[23] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=61
[24] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=62
[25] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=71
[26] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=72
[27] http://thedataweb.rm.census.gov/TheDataWeb_HotReport2/econsnapshot/2012/snapshot.hrml?NAICS=81

Three components of travel journals: spatial, temporal, human agent

The first travel journal properties to discuss are spatial, in the form of toponyms and POI names that are used within the journal. The spatial component includes any names that can be georeferenced to define the specific or proximal location of the respondent. Specific locations are areas visited or observed by the respondent. Proximal locations are geographical settings observed or mentioned by the respondent but not actually visited. These spatial travel journal components, in the form of toponyms, have a structure and hierarchy that can be used in georeferencing. The second travel journal properties are temporal elements exhibited as time that is explicitly stated or implied. Temporal elements provide the time range of the POI name's existence, the age of the human agent, the time range of the overall trip, and the length of time the trip took from origination to destination. Temporal elements in the context of analysis, also include temporal elements and consistency between the gazetteer resource and the time period as the travel journal. The third travel journal property or component is the respondent or agent, mainly human, and often the author of the travel journal, who transcribes his or her own observations or experiences. The human agent will guide the readers on their travel and provide some presupposition to travel origination and purpose. The availability of the three components contained within a travel journal formulate a framework that will encapsulate a geographic instance or a place visited or observed at a specific time. All components are required to provide the context to a geographic instance and minimize computational disorientation.

A human agent is useful for describing a geographic instance of a place. The concept of space is abstract and often a direct manifestation of the human agent and his or her surroundings. These events are global, local, personal, and major or minor in magnitude. Examples of an event are the author visiting his grandparent's house in 1989; or marines raising of the American flag at Iwo Jima on February 23, 1945. An event is based on human action, and these examples represent a geographic instance based on human events. Understanding how an author guides his or her readers within a travel journal provides a structure within the GIR process to position a place within a defined focus.

Travel journals provide a literary map or guide denoting and describing places visited or of interest to its author. Travel journals are subjective attempts in orienting readers and navigating them through a personal journey. Eric Bulson (2006) states that novels produce space and that readers both consume and produce space. Pritchard and Morgan (2000) recognize space and place as socio-cultural constructions rather than simply as physical locations. People and their perceptions dictate how they will view and formulate their travels. The socio-cultural construction of a place is important, as the construction can be implied within a travel journal. The journal produces space, but it is left to the reader to interpret the space. The interpretation of space by the reader can be the same as or different from the author. Bulson notes that an abstract level of a sense of location can be found in literary maps and can take on real or fictional forms. The geographic instance is dictated by the socio-cultural biases of the author, which is further convoluted by the socio-cultural biases of the reader. Places are subject to human

interpretations, and subjectivisms impact how the authors and readers describe a place. Toponyms and POI names must maintain some sense of universal recognition to avoid disorientating the readers completely, and in retrospect the application to process such data. A primary supposition of the research contained in this dissertation (which is not a supposition of Bulson's work) is that the geographical settings and places described by the author are real. This abstract reasoning of space provides a challenge for defining the geographic instance through computational means that requires instructions on how to process data.

In Bill Bryson's travel journal the author repeatedly described places using descriptive names that best fit his perception of that place. Names like Fungus City, Fudd County, Dunceville, or Dog Water are fictional names created by the author, but used to portray the author's impression of locations visited. Fictional names and other subjective determinants cause a high degree of erraticism during computational identification. They are based on the author's impression of the town and not the actual name of that town, although some jargon-based, subjective naming has been incorporated into gazetteers used with VGI and crowdsourced data (Rice et al. 2011, 2012). In the GIR process, placenames specific to narrow socio-cultural groups can cause vagueness and geo/geo and geo/non-geo ambiguities to both readers and computational algorithms, and the result is omission of the places the author is describing.

This study is not concerned with locations that are fictional, contradictory, unconventional, or imaginary, but is focused on names recognized by a valid reference document. Names alone, even those referenced by gazetteers, do not denote a place.

Computational indeterminacy is very high due to geo/geo and geo/non-geo semantic ambiguities. To properly add some sense of orientation and pragmatic association to a location defined by the literary mapping of a travel journal, a collection of space, time, and human agent must be known, the place must have a known name, and that name must exist in a gazetteer or other placename reference.

Human Factors

Yoel (1992) wrote that tourist destinations are sometimes unknown at the time decisions are being made. It is difficult to know where people will travel, as their destination may or may not be known due to the dynamics of decision making. What is known is that human attitude is a determinant to a decision to stop at a specific destination. Human attitude is a decisive factor in travel destination behavioral choices (Um and Crompton, 1990) and the knowledge of human attitudes towards a trip provides assumptions to determine the mode of transportation, possible destinations, and purpose of the trip. The presupposition elements based on human attitudes denote a subjective belief that provide specific data for origination and destination choices.

A human factor contributing towards developing a presupposition data repository are past experiences. Sonmez and Graefe (1998) described past experiences as an influential motivator to selected travel destinations based on associated risks of the destination. Prior destinations provide a preliminary foundation that decide where people may travel, the purpose, its personal spatial-human relationships, frequency of the trips,

and time and distance of the trips. Subjective reasoning defines "push and pull" factors that describes the reason a destination was selected.

Bill Bryson's travel journal accounted for push and pull factors that determine his destinations. He traveled to Winfield, Iowa to reminisce about his grandparents' town; New Salem, Illinois, Hannibal, Missouri, and Warm Springs, Georgia to visit historical sites; Pella, Iowa to see the Dutch community; Santa Fe to visit a niece; northern California to Sequoia and Yosemite National Parks; and Cooperstown, New York to see the Baseball Hall of Fame. The travel journal also includes places mentioned as a possibility to visit, but which were not destinations such as Minneapolis to see the Twins play baseball, since the game was played in Baltimore or Redwood National Park, which was too far from the author's current destination at Sequoia National Park. A "push and pull" factor determines what drives the tourist to participate in a trip and what pulls them to a specific spot. The "push factor" is motivation that drives the tourist to decide on a trip, such as boredom, health issues, realizations, or businesses. Destination choice decisions are based on "pull factors" that are tangible characteristics pulling tourists towards the destination (Jonsson and Devonish 2008). These pull factors can refer to a destination that is attractive to potential visitors, including historical and cultural resources, beaches, and personal visits to family. Push and pull factors are highly identifiable in travel journals where personal thoughts and reasoning about a destination are documented. These factors can help explain why a choice was made for a specific trip.

Other factors such as gender and age have impact on travel destinations. Examples of gender impact were described by Andreu et al. (2005), who asserted that females have a stronger motivation to travel than males. They also found significant gender differences regarding travel motivations, with male tourists preferring more recreation and activity in the destination, and female tourists having stronger relaxation and escape-based motives. Swain (1995) further defines gender involvement with travel and tourism as a system of culturally constructed identities, expressed in ideologies of masculinity and femininity, interacting with socially structured relationships in divisions of labor and leisure, sexuality, and power between women and men. Gender has been shown to have some determinable purpose for a specific trip, but age can also provide reasons for a trip as well.

Jonsson and Devonish (2008) believed that the age of a tourist has a significant effect on cultural and relaxation-based motivations. Post hoc tests reveal that tourists in the oldest age category (56 years and over) report significantly stronger cultural motivations while the 36 to 55 age group had significantly stronger relaxation-based motivations. This dissertation does not attempt to test theses about the influence of age or gender on distinctive roles in travel behaviors, but this study will explore the author's characteristics and if those characteristics can improve georeferencing techniques.

Bill Bryson is a Caucasian male who was in his mid-30s when he traveled across the US. Bryson did exhibit recreational and activity characteristics attributed to males, but Bryson had a significant cultural motivation which is pronounced during his travels through the Deep South. Bryson's travel through the Deep South described in his travel

journal was his second encounter. His first experience of the Deep South occurred while Bryson was in college and he and his college friends had an uncomfortable encounter with the local inhabitants in Georgia. The local inhabitants looked warily on Bryson and his friends that made them uncomfortable. Bryson remembered three freedom riders (two white and one black man in their early 20s) being murdered five years prior in Mississippi, and no one was charged for the crime. The murder of the three freedom riders embodied the Deep South for Bryson, and his views of the Deep South are reflected in his travel journal.

From the time Bryson entered Tennessee he describes it as being in another country. While describing people in the Deep South Bryson exaggerated the way people spoke such as the word "right" being pronounced "rat", "square" as "skwaya" or "breakfast" as "breast". Bryson critiques how a waitress pronounced "Can I help you" as "Kin I hep yew". When Bryson entered Mississippi, he exaggerated the welcome sign based on personal experience as "Welcome to Mississippi. We Shoot to Kill." Bryson's experience from his college days had such a negative portrayal of the Deep South that when he visited Tuskegee, Alabama, which was predominantly black and dilapidated and later Auburn, 20 miles northeast of Tuskegee, which was predominantly white, clean, and wealthy Bryson criticized the disparity in wealth. There was one positive experience during Bryson's travel through the Deep South: while in Tupelo, Mississippi, a larger town with a strip mall, hotels and restaurants, the author exhibited a sense of relief. Bryson welcomed the convenience of being in a clean and comfortable place that was reasonably priced.

The travel journal used in this study has presupposition elements required to develop information about trip originations, destinations, and mode of transportation to georeference the toponyms and POI names. If the identity and naming of a spatial area are defined by socio-cultural determinants than it stands to reason that space and its names are dependent on human factors at a given temporal stage. Personal attitudes, previous experiences, environmental relationships, age, and gender all formulate a specific behavior towards travel. These travel behaviors provide presupposition knowledge used to prepare the GIR workflow based on what is known about the human agent. This framework, developed from relevant previous research, will be re-examined in the following chapter on methodology.

User Contributions to Geographic Information and Gazetteers

The advancement in network technology and its inclusion of social interactions popularized the concept of VGI (Goodchild 2005, 2007) where geographic data are provided freely by the public on a platform designed to capture or encourage data input. As technology advances, the temporal component encourages and supports real-time communication among its participants (Bargh and McKenna, 2004). VGI's promotion of geographical and temporally-sensitive data made possible by technological advancement provides an opportunity to collect additional geographic representations within a more restricted temporal framework. VGI is important to the GIR process as it is the collection of data motivated by a human agent within a narrow temporal focus that can assist in

narrowing the identity or formulating a conceptual framework of a toponym or POI name.

Motivations drive users to voluntarily contribute geographic information on a public forum or through public geo-media. Coleman et al. (2009) list eight constructive motivational factors that include altruism, professional and personal interest, intellectual stimulation, protection of personal investment, social rewards, personal reputations, self-expressions, and pride of place. Coleman et al. (2009) also list three negative motivational factors as mischief, agenda, and malice. Also, motivations to voluntarily provide geographic data are influenced by the quality of social media to assist individuals about a social movement (Hertel et al. 2003). Additionally, motivations increase when the author and contributors both have a common purpose and common interest; in other words, people are more driven to participate if their interest matches the author's. In the realm of VGI, the author's agenda should generate enough interests to provide motivation so that people will promote the availability of data and store the data that is accessible later for retrieval.

Voluntary motives reflected by VGI are internal forces that provide the energy to move people to contribute based on personal interests and technical conveniences. The motivation of self-promotion suggests that some people contribute to be rewarded for monetary gains, competitive advantages, or individual gratification (Goodchild 2007). Community concerns associated with political or environmental issues can often encourage user's contributions in the hopes of assisting or aiding others (Conrad and Hilchey 2011). During times of crisis, the public turns to social media networking sites

to learn and communicate (Verma et al. 2011). Off-line social movements such as civil rights, labor, or peace movements can, by definition, motivate people collectively to solve a common problem (Hertel et al. 2003). A perception that contributions will improve an individual's job performance within an organizational context, can increase participation and additional motivation (Teo et al. 1999). Finally, the ease of use and personal enjoyment when blogging or surfing the web are important factors influencing motivation (Hsu and Lin 2008; Teo et al. 1999). Self-esteem, community concern, social movements, and enjoyment are several factors contributing to the motivation of users to voluntarily contribute to social media or forums. Motives drive the users to contribute geographic data, and in return the contributions allow a forum for those users to express their thoughts.

Social media technology platforms provide an online presence that allows people to genuinely reflect their thoughts (Yarkoni 2010) and personalities (Hirsh and Peterson 2009), and create what are described as self-narratives. The extension of personality traits and thoughts along an online medium, including mobile devices, will provide the opportunity for data input on a continual basis. Technological advancements — including hardware, software, data, and networks — allow the input of data about user activities into social media applications. These activities include accommodations for people with vision impairments to navigate specific routes avoiding difficult obstacles, with information generated by geocrowdsourcing (Rice et al. 2011, 2012a) and geosocial networking (Waters 2013). Other user activities include participation in the social feminist movement (Elwood 2008); earthquake disaster relief (Zook et al. 2012); and

urban planning design and development (Seeger 2008). VGI has shown importance to society as an informative means to relay knowledge of a geographic occurrence, but technology advancement is required to provide a platform to communicate and express those thoughts on a real-time basis. The increase in use of VGI can promote personal reflections describing an event of importance at a location close to or in real-time.

VGI and gazetteer-based geoparsing has been a source for identifying navigation obstacles by Rice et al. who document their methods in several comprehensive technical reports (2012b, 2013b, 2014, 2015); and several peer-reviewed articles (2011, 2012a, 2013a, 2016). They outline the nature of geocrowdsourced information and how placenames and text-based geographic descriptions can be used to provide information to the disabled. Qin et al. (2015) and Aburizaiza et al. (2016) extend this work through quality assessment and spatial footprints generated through geoparsing and spatial logic.

Georeferencing and Gazetteers

Data management of placenames and their temporal information contained within gazetteers are important in providing a knowledge-based repository of a placename. Named Entity Recognition (NER) aims to classify entities as persons, organizations, dates, or products, in addition to geographic entities (Smith and Crane 2001). The attributes included within a gazetteer provide an ideal resource for name matching of geographic toponyms and obtaining the toponym's geographic coordinates. Since specific names are being identified, gazetteers are highly involved within the name matching process (Hill 2006). A list of geographic toponyms aids the NER process, but

several challenges remain, including word or phrase order variation, derivation, inflection, synonymy, and homographs (Nadkarni et al. 2011). NER is a metadata approach to tag and resolve placenames which requires the aid of a gazetteer to identify the reference of the placenames.

Placenames at a given time are widely used in conversation, correspondence, reporting, and documentation (Hill 2006). These places, parsed from documents of utterances, can be referenced to the names contained within a gazetteer identifying their coordinates. Information contained within gazetteers is essential in obtaining the toponyms required to define a primary state name and narrowing the location of a toponym or POI name. Current gazetteers provide a comprehensive list of spatially indexed geographical information (Smith and Frew, 1995). Data management of placenames and their temporal information are important knowledge-based repositories to georeference a placename to their respective mapping location. Traditional gazetteers are highly valued for managing information for political toponyms and larger institutions, but the management of POI names in gazetteers is limited. The limited use of POI names within a gazetteer can be remediated by heuristics and web locators.

Traditional gazetteers are not focused on local POI names, but a heuristic approach can assist in the identification and disambiguation of POI names. The heuristic task includes basic linguistics name recognition techniques such as sentence boundary detections, and morphological detection of text semantics (Nadkarni et al. 2011). These tasks recognize specific syntax triggers that can encapsulate a word (prepositions), identify cities or states (“city of”, “state of”), or associate a proper noun with the use of a

definite article (The museum of). The heuristic approach governs a rule-based functionality and encapsulation of geographic texts often not found within a gazetteer. Despite its advantages the heuristic approach is limited to those POI names containing annotations such as named entity recognizer or part-of-speech to initiate a computational read to identify the text as a geographic name.

Georeferencing POI names using online mapping locators such as Google Map or Nominatim can aid in the recognitions of the POI names. Subscribers pay to post their business names onto a map to allow potential customers to locate their businesses. An article by the Wall Street Journal[28] reported that mobile advertisements associated with maps accounted for 25 percent of the estimated $2.5 billion spent on the overall mobile advertisements. This monetary assessment demonstrates the popularity of mapping applications (especially for mobile devices) that can drive business and organizations to subscribe to a proprietary mapping service. As more POI names are added to online mapping resources, the ability to capture those names becomes greater. By combining the use of online maps, heuristics, and metadata approaches, most available POI names can be identified and placed from a textual narrative.

Other gazetteer-like references for POI names are available but lack historical attributes. OpenStreetMaps[29] allows their users to request changes to their maps, but nothing in their procedures require managing historical information for a site. Garmin, a GPS device maker, allows third-party databases to upload to their product, but the main

[28] Vascellaro, J. E., and Efrati, A. (2012, June 4). Apple and Google Expand Their Battle to Mobile Maps. Wall Street Journal. Retrieved from http://online.wsj.com/article/SB10001424052702304543904577398502695522974.html.
[29] http://wiki.openstreetmap.org/wiki/Points_of_interest

components required for data compatibility include POI names, coordinates, and speed limits, but excludes temporal data. Google Maps lists available field names for their APIs,[30] but indicates no temporal information except for time and durations of events (concerts, fairs, or private parties). Names and coordinates are common fields contained in both Google Maps and OpenStreetMaps, but very little capacity exists for temporal data. Temporal information for a POI name is not currently managed by popular mapping services, and given historical evidence about the longevity of a POI name for a given location, the name of a specific space will likely change over time.

Temporal elements are strongly encouraged within gazetteers to represent changes (Hill 2006) and answer when an event took place (Gey et al. 2010). Temporal information is either explicit (temporal expressions) or implicit (metadata, and ordinal text) (Peregrino et al. 2012). Peregrino et al. defined explicit temporal data as actual time and date stamps that can be retrieved within the pages of the documents, forum's comment headers, hash tags, or web URL. Implicitly, temporal data can be retrieved by metadata if a timestamp of that toponym exists within the gazetteer. Explicit temporal data is advantageous for social media, news articles, or micro-blogging web sites where specific date and time states exactly when an event occurred or a narrative was written. Implicit temporal data is advantageous for chronological ordering where date and time are not used but contained within a metadata resource or described as an ordinal narrative.

[30] https://developers.google.com/maps/documentation/geocoding/?csw=1#JSON

VGI recognizes the value of user's contributions at a temporal instance and its participatory role to aggregate information. VGI shows strength in dynamic situations where explicit temporal items are of importance during emergency situations like forest fires (Kebler et al. 2009) or micro-blogging which provides a real-time rate of people's reaction and opinion (Stefanidis et al. 2013). When temporal data is provided explicitly, an assumption can be made placing the placenames to the time identified; when no time is provided an indirect or implicit means of generalizing the temporal analysis within a VGI platform is required based on the local geographic entities identified within the documents. Gazetteers with temporal components can identify the existence of a toponym and answer questions based on current, historical, or recent events. Temporal data for this study will be restricted to the year the travel journal was published and the existence of data contained by current gazetteers.

Travel journals and social blogs are becoming common within dynamic platforms (Schmallegger and Carson, 2008) that allow real-time entries of geographic-sensitive data. Travel journals and gazetteers demonstrate how human agents can implement a NER process to identify, name, and label a spatial area. GIR applications utilize different approaches for different types of documents, but these applications are dependent on the existence of a gazetteer to provide the coordinates for a toponym identified and matched to a name contained within a gazetteer.

Temporal Disassociations

The longevity of toponyms from the time Bill Bryson's travel journal was published (1989) to the current time this dissertation (2016) creates a risk of missing placenames that existed in the time the journal was published but not when the study was conducted. A basic analysis was conducted between 1989 and 2016 to determine if any drastic changes to the taxonomic ordering occurred at a higher-level scale using country names. In the 27 years, it was found that seven major global occurrences resulted in name changes at a country level. Within the United States one change was made at a county level when Dade County, Florida changed its name to Miami-Dade County in 1999. All toponyms and local placenames contained within Miami-Dade County inherited the change in the county name. Recognizing the temporal sensitivities of human-defined politicized toponym boundaries cautions the use of gazetteer resources that have been created prior to any toponym name changes.

Radding and Western (2010) noted the long-term stability of people's home-land as assurance for the long-term viability of toponyms. Also, the stability of larger political boundaries defining states and countries provide long-term viability of toponyms contained within a gazetteer. A preliminary research question posed at the start of this work was, "How many high-level toponyms whose names or boundaries changed between the year this study was conducted (year: 2016) and the year the travel journal was published (year: 1989)?" It was found that between the 27 years that elapsed from the time of Bryson's travel journal was published to the start of this dissertation some major disruptions have caused significant political ramifications impacting the names of

higher-level toponyms and its subsequent toponyms. Figure 7 summarizes the major political events during the 27 years that led to significant name changes.

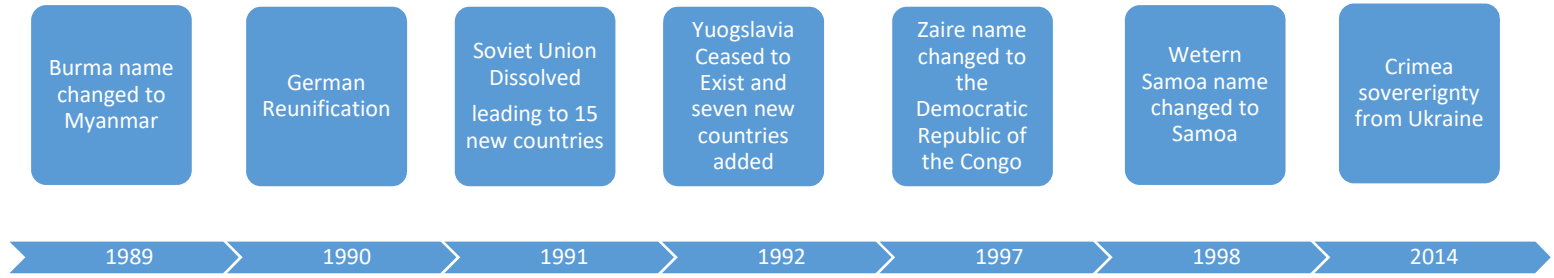

Figure 7 Between 1989 when Bryson's travel journal was published to 2016 seven countries either underwent major transformation or dissolved leading to the creation of new countries or name changes

The last 27 years experienced high-level disruptions that caused not only countries to add or remove names, but disrupt the toponym associations of subsequent toponyms and POI names. The land that formerly contained Yugoslavia is as an example of an area with different political name ownership at different temporal instances (Smith and Crane 2001). Before 1992 Yugoslavia existed as its own country, but during and after 1992 it was divided into seven different countries. Another example of a disrupted political boundary is Crimea whose annexation by Russia from Ukraine has been contested and is not universally recognized.

Contesting claims (Lay et al. 2010) over political boundaries are a concern as they confuse political sovereignty over a specified space. Gazetteers are very susceptible to political differences. The concern over providing factual information of a place location and ownership is important as various states claiming the same location offer different, conflicting geographic boundaries. Information contained within maps or gazetteers is dependent on political governance. Geographic representations commonly reflect their popular political context, and should a political disruption occur, gazetteers will reflect those changes within its proper political governance. The US governing body that records authoritative placenames is The U.S. Board on Geographic Names.

Implicit temporal data derived from gazetteers and locator applications will be the primary means to determine the temporal relevancies or the local POI name. The travel journal used for this research was written and published before the creation or popularization of electronic metadata for local POI names. Toponyms and POI names

are temporally sensitive and the events that occur at both global and local levels can alter the definition of a spatial area.

Geographic Information Retrieval

The GIR process identifies and indexes geographic content (Larson 1996) by retrieving, disambiguating, ranking, and storing toponyms and other placenames from textual narratives located in a multimedia platform. The indexing stage is a repository of ranked toponyms with geographic information that are stored for future requests and spatial queries. The GIR process extends the definition of toponyms from a purely narrative description into a quantifiable and georeferenced data set. The identity and indexing mechanisms emphasize improving the quality of GIR with focus on unstructured documents (Jones and Purves 2008). Gazetteer internal structures are developed as an elaborate hierarchy-tree list of objects where placenames reside within a primary data table and its subsequent elements, codes, relationship-types, and schemas are linked to that primary data. Gazetteers must be managed to improve the evolving list of the static "geographic-logic" file structure to adapt to its proper temporal sources. This geographic-logic file development shifts the focus of a gazetteer from a static document to an organic metadata resource for geographically aware and temporally sensitive search technology.

A demand exists to develop geographically aware search technology that can index and retrieve web documents according to their geographical context (Vaid et al. 2005). Data contained in electronic documents such as news articles and social media are often fuzzy in nature (Jones et al. 2002; Silva et al. 2006), but the information contained within unstructured text is more numerous than its structured counterparts (Rauch et al.

2003). Collecting additional geographic data to improve existing metadata resources is an opportunity to evolve gazetteers into a more formidable data resource.

To improve the identity and indexing of placenames, the GIR process relies on actual geographic terminologies to emphasize the proximity, containment, and spatial operations to place a toponym in its correct location within an appropriate measure of confidence (Larson 2011; Andogah et al. 2012). This dissertation expands upon the GIR process and implements an outline of the requirements that will effectively confirm the identity of a POI name and toponym, and place those names based on gazetteer and locator references. Peer-reviewed applications made important contributions to the overall GIR process and are listed in Table 8. The Georeferenced Information Processing System (GIPSY) (Woodruff and Plaunt 1994) visually portrays the likelihood of an area's location by polygon overlays. The Perseus Project (Smith and Crane 2001) and Edinburgh Geoparser (Grover et al. 2010) focus on the geo-parsing of historical data. Spatially Aware Information Retrieval on the Internet (SPIRIT) (Silva et al. 2006) addresses the issue of alternative names of a toponym resulting from abbreviations or spelling variations. MetaCarta (Ruach et al. 2003) tries to imitate human processes using heuristics and data mining. Finally, Web-A-Where (Amitay et al. 2004) emphasizes the importance of a toponym's hierarchy. The applications in these previous studies approach the GIR process by formulating a conceptual design to geoparse and georeference toponyms contained within a digitized textual narrative.

Table 8 List of names parsing applications

Application Name	Description	Source
Georeferenced Information Processing System (GIPSY)	Visually displays weighted overlaying polygon relationships (synonymy, kind-of, part-of) Relationships based on lat/long of toponyms found Polygons with highest display most likely represent the place of occurrence	Woodruff and Plaunt 1994
Perseus Project	Identification and categorization of historical names and disambiguation of name classes	Smith and Crane 2001
MetaCarta	Commercial application that mimic human process using heuristics and data mining to extract relevant data from unstructured text	Ruach et al. 2003
Nominator	A recognition system of known names and discovery of new names through a machine learning technique that limits the use of an authoritative database	Wacholder et al. 1997
Web-A-Where	Use of NER and data mining approach to identify all geographic entities, assign a geographic location, assign a confidence level, and derive a focus for the page.	Amitay et al. 2004
Spatially-Aware Information Retrieval on the Internet (SPIRIT)	Emphasis on the ability to recognize alternative names of the same place and places similar or of proximity to the location in the query	Silva et al. 2006
Edinburgh Geoparser	Focuses on a specific group of collections within a narrow geographic scope within the United Kingdom for specific toponyms based on lexicon, attributes, and lexicon variation components.	Grover et al. 2010
MyTravel	Provides a social application that integrates a geographic component to display a user's current or future travel routes	Cestra et al. 2011

Scientific studies exist that involve geo-parsing and georeferencing points-of-interest names. A major difficulty associated with geo-parsing and georeferencing local-level names representing neighborhoods, streets, and human-developments is that these names do not appear in gazetteers (Gelernter and Mushegian, 2011). The omissions of local-level names from gazetteers renders an NER processes useless as they require metadata resources for name matching. A means to alleviate the lack of local POI names contained within gazetteers is to link the POI name to the toponyms whose names can be georeferenced (Drymonas et al 2011) or using current geocoding tools that contain local entities, such as Google.

POI names are settings that relate to a smaller geographic footprint on a map representing a location recognizable by the person searching for that entity. POI names are important because they denote settings such as churches, restaurants, community centers, residencies, or work-places which are designed for humans to conduct specific activities. Because POI names represent an actual place with a smaller geographic footprint, they are often contained within or intersected with the toponym's boundaries. The spatial association of physical developments representing a point-of-interest allows a taxonomy relationship between POI development and toponym boundary. Affixing POI names to a toponym taxonomy tree improves taxonomic resolution. The capability for identifying a human setting contained within the selected states permits a more detailed geographic analysis within the GIR process.

Geographic scope is a region or area for which the document is geographically relevant (Andogah et al. 2012). Geographic scopes are authoritative toponyms such as

districts, counties, states, or nations which can provide geographic containment or proximity relationships for POI name that are geographically portrayed as a point feature from a commercial application. There can be one or many geographic scopes per document (Amitay et al. 2004), but every document will have at least one geographic scope. POI names currently lack a formal metadata resource for georeferencing, but a presupposition can be made that a POI name will always have a spatial (containment, proximity by distance) relationship to a specific geographic scope for that document. The association of a POI name to a geographic scope and all toponyms contained within that scope improves the resolution of a taxonomy tree.

A geographic scope is required to contain the toponyms and POI names contained within a gazetteer for each chapter of the travel journal. For this dissertation, the geographic scope are the state names relevant to the section of the travel journal. The course resolution of the state names contains the subsequent toponyms which is used to establish a toponym hierarchy tree. The list of state names can minimize the ambiguities for local POI names based on the geographic locations. Relationships of geographic entities have “adjacent-to”, “part-of”, “toponym frequency” or “importance by population or size” relationships (Andogah et al. 2012, Silva et al. 2006, Moncla et al. 2014). Within a document, the name "Paris" could indicate a geo/geo or geo/non-geo ambiguity, because of additional instances of the same name. A state name or georeferenced POI name can affirm the characteristics of the instance of Paris and resolve ambiguities. Geographic names that represent a country containing Paris (i.e. France), cities within

proximity to Paris (Marseille or Lyon), or listing Paris as a major city implied that the "Paris" contained within the text document is geographic in nature and a place in France.

The taxonomy tree is a collective and ranking approach to build a geographic scope that contains each location and disambiguate a POI name (Andogah et al. 2012). The taxonomy tree collects a group of toponym nodes and arranges them in a cascading political hierarchy order [Country name/State/City/District/POI]. The leading node provides a check confirming the subordinate nodes. The tree takes advantage of the more stable, long-term availability of political toponyms whose boundaries partially or fully contain the POI development. By allowing the POI name to terminate the toponym tree it defines the toponym resolution at the smallest geographic footprint based on a placename.

The geographic scope is important in its hierarchy schematics and heuristics. A hierarchy as seen with the toponym's taxonomy tree can be highly developed when all or most pertinent geographic information are available (Ding et al. 2000, Amitay et al. 2004). It is through the hierarchy system that the toponym resolution is developed and a location analysis is derived for a given POI name. Toponym resolution is a process of assigning a placename identified in a text to a single non-ambiguous place on the earth surfaces (Buscaldi 2011, Andogah et al., 2012). In the case of this dissertation, the geographic scope is defined as the primary state name pertaining to each section of the travel journal and its adjacent state names. By providing and associating the toponyms in groups of nodes for a single POI name the toponym's resolution can improve.

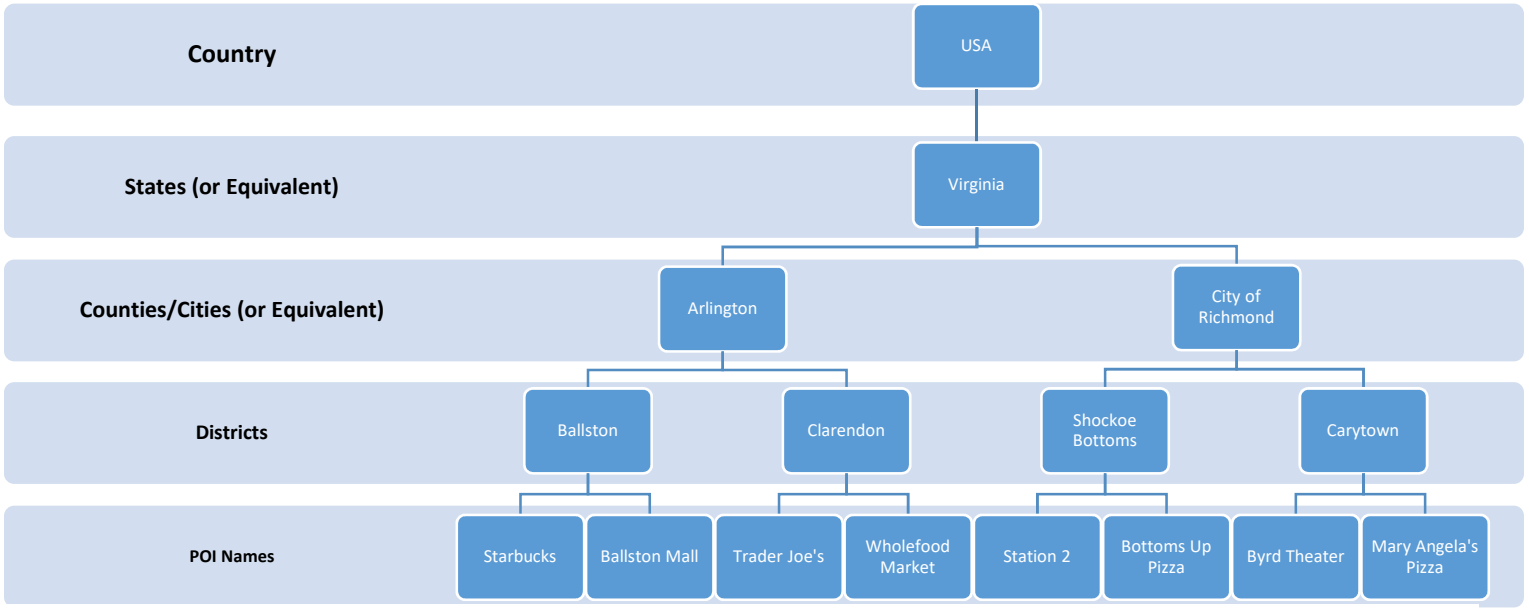

Figure 8 The geographic hierarchy within the United States lists all subsequent abstract political boundaries terminating at a physical human development level.

Natural Language Processing

NLP is a complex operation as it provides machine-learning techniques to recognize names based on their structure and content (Amitay et al. 2004). The tasks include sentence boundary detection, tokenization, and morphological detection of text semantics (Nadkarni et al. 2011). These tasks recognize specific syntax triggers that can encapsulate a word (prepositions), personal names (Mr. Mrs. Dr.), cities ("City of"), or a definite article ("The museum of"). The integration of NLP concepts, especially within GIS technology, that recognize textural spatial operations and geographic terminologies can provide answers to user-based questions. (Cali et al. 2011). NLP governs the rules for machine-learning functionalities and encapsulation of texts which contains candidate geographic names used by this dissertation.

A heuristic approach can assist in the identification and disambiguation of POI names. Combining non-traditional geospatial databases with spatial imagery can identify geographical places and update or upload to a gazetteer (Michalowski and Knoblock, 2005). The heuristic tasks include basic linguistic name recognition techniques such as sentence boundary detection, and morphological detection of text semantics (Nadkarni et al. 2011). The heuristic approach governs a rule-based functionality and encapsulation of geographic texts often not found within a gazetteer. Despite its advantages, the heuristic approach is limited to those POI names containing annotations to initiate a computational read, in which case a named entity recognition system is preferable. For this study, the NLP process will be used to identify and tokenize potential text representing locations. Personal names and temporal elements parsed by the NLP application will be searched

but not used. The primary importance of the NLP application will be focused on categorizing text that are locations.

Vagueness and Pragmatic Halo

Natural language is large, unrestrictive, and contains ambiguities that can lead to semantic problems in understanding its meaning (Nadkarni et al. 2011). The complexity of natural language is apparent in data retrieval where spelling variations, errors, and geographic/non-geographic (geo/non-geo) ambiguity pose difficulties in geoparsing (Wang et al. 2005). To remedy the difficulties and associate meanings to disambiguate texts from information retrieval, an NLP and named entity recognition (NER) were developed to automate the comprehension and meaning of text retrieved from documents.

Gervais et al (2009) stated that uncertainties will almost always exist but can be minimized by gathering required data such as position, time, and theme for geospatial data quality. Position, time, and theme correlates to the three-components identified for the conceptual framework of spatial, temporal, and human agent, and reflects the "atomic" view of geographic information espoused by Longley et al. (2015) which includes location (x,y), time, and attribute. The requirements imply that geo-locating the location of toponyms or POI name from the single statement alone is difficult without formulating a context from the documents. Human tendencies to state location within a pragmatic halo (Lasersohn 1999), and vagueness of a specific location (Varzi 2001). Location vagueness generalizes a location of an event rather than being specific. Varzi (2001) provides examples of vagueness such as Mt. Everest, downtown Manhattan, and

country boundaries within Lake Constance. The three examples Varzi provide are fuzzy areas whose locations are only known as "part-of" an area described rather than a specific location. A pragmatic halo is the utterance of a location not yet arrived at by the respondent, but within an acceptable distance such that the utterance is consider true. People can declare that they are at a specific location when the reality shows that they are in proximity to the mentioned location. A submitter to a message states they are at a store when they are at the store's parking lot as denoted by the geo-tagged submission from the mobile device. Pragmatic halos imply false statements as true and computational analysis should consider such statements as true while recognizing the reality of space and time. People accept the truthfulness of location that is described vaguely or in proximity to a destination, but at a computational level informal and imprecise communication is taken as literal. This dissertation will not attempt to eliminate vague descriptions or locate the actual position of the author, but will focus on extrapolating toponyms and POI names and correctly georeferencing them.

Ambiguities

What this dissertation will do is minimize semantic ambiguities to correctly label a text as a placename and georeference the text to its proper location. Toponyms and POI names by themselves are insufficient determinants of the site's true location. Names often suffer from structural and semantic ambiguities (Wacholder et al. 1997). Semantic ambiguities such as duplication, geo/non-geo, and geo/geo, structural and semantics, spelling variations, name changes, and lack of uniqueness exist among varying spatial

locations either in distance or proximity of one another (Zhang et al. 2012; Wang et al. 2005; Grover et al. 2010; Woodruff, and Flaunt 1994; Wacholder et al. 1997). This research confirms the importance of accommodating and resolving problems associated with spelling variation and alternative names, but most focus is placed upon proper naming conventions for placenames that will be referenced to a state name to minimize semantic and structural ambiguities. Cave (2016) addressed several of these issues in the process of georeferencing a historical, fictional travel narrative based on real geographic locations, where alternative spellings appeared as a byproduct of language translation and historical changes in toponym spelling.

Sperber and Wilson (1985) stated that a speaker wants to convey a single atomic proposition, but the complexity of human thought is made up of atomic thoughts. Conveying a single meaning using a specific word that can yield different interpretations is a struggle for GIR studies. Human communication is often a platform of fuzzy and vague statements, allusion, metaphor, and not of actual literalness. Semantic ambiguities and vagueness often lead to computational confusion about the validity of a word that can potentially relate to a geographic name. Ambiguities such as duplication, geo/non-geo, and geo/geo, is the major challenge that exists for any GIR study and exists among varying spatial locations either in distance or proximity to one another (Zhang et al. 2012).

The peer-reviewed studies (Table 9) are focused on removing or minimizing the impact caused by ambiguities to provide the correct location of a placename. Disambiguation improves the quality of the data by narrowing the meaning of a

geographic context to a single location reference. The disambiguation of geographic placenames is required to georeference the placename, and minimizing geo/geo and geo/non-geo ambiguities is a challenging obstacle experienced by GIR applications (Gelernter and Balaji 2013, Amitay et al. 2004, Woodruff and Plaunt 1994, Smith and Crane 2001, Silva et al. 2006, Cestra et al. 2001).

Table 9 GIR applications means to minimize geo/geo and geo/non-geo ambiguities.

Application Name	Ambiguity Resolutions	Source
Georeferenced Information Processing System (GIPSY)	Polygon overlays of potential geographic candidates with the highest overlay most likely representing the place of occurrence.	Woodruff and Plaunt 1994
Perseus Project	Disambiguate geographic names based on identification and categorization.	Smith and Crane 2001
MetaCarta	Measures the confidence that a reference name (n) refers to a point (p) based on georelevance.	Ruach et al. 2003
Nominator	Only words listed in the gazetteers are considered. Also, assumed a "Single Sense per discourse" principle. Where an ambiguous term is likely to mean only one of its senses when used multiple times unless specifically qualified (i.e. He drove from Portland, ME to Portland, OR)	Wacholder et al. 1997
Web-A-Where	Apply confidence weight based on State abbreviation, geographical entity with significant population size, and multiple occurrences of same name. Unresolved names are given less weight to minimize its impact.	Amitay et al. 2004
Edinburgh Geoparser	Evaluate "Type:Token Ratios" for the entities in the collections to discover whether there are differences in lexical variation. The ratio determines if a document has a higher count of people or location names.	Grover et al. 2010
MyTravel	The focus is on geographic names that can be located on a map. A location that can be mapped restricts ambiguity issues found within a text document.	Cestra et al. 2011

The studies reviewed for this dissertation minimize the impact of fuzzy and ambiguous statements by providing heuristic reasoning, named entity recognition, gazetteer associations, and weighted confidence by category, or spatial overlay. This study attempts to extend the GIR process to include a combination of frequency and clustering analysis to minimize ambiguities and place the toponyms and POI names found within a document to their proper location. The need to identify all geographic candidates, and a common hierarchy requires the use of NLP and NER processes, using a gazetteer and on-line mapping locator as a reference source. The three resources, gazetteers, heuristics, and on-line mapping applications are designed to identify names, alleviate ambiguities, and provide geographic information for toponyms and POI names. The final deliverable is a list of all toponyms visited or observed by the author of a travel journal and a map showing the order in which they were visited (Figure 6).

Conclusion

The GIR process parses through documents and utterances to index spatial data to their correct geographic location. As more documents and utterances are collected, GIR is becoming prevalent in referencing geographic locations. A major challenge experienced when georeferencing toponyms and POI names is semantic ambiguity. Geo/geo and geo/non-geo ambiguity complicates automatic determination if the data collected is geographic in nature, and if it is geographic data, the correct spatial location for the geographic data. As technology evolves and continues to provide a platform for social interactions and personal contributions, the geographic data are associated with an

individual or groups of individuals fueled by motivational factors to provide geographic data that are temporally sensitive. A human and temporal component can narrow the context of the geographic data by knowledge of locations of interests by the individual and the temporal existence of the geographic location. The travel journal used for this study will rely on temporally sensitive gazetteers and POI locators to identify and reference placenames that existed in the past but not now. The travel journal can make use of toponyms that are of interest to the author based on residential, employment, and family history. As more information about the time of the event and author are gathered, a presupposition can be made of possible places the author may have visited in the travel journal.

CONCEPTUAL FRAMEWORK

The conceptual framework for this dissertation is based on the three components introduced earlier, which describe an instance attributed to a geographic-related event. A geographic instance is an action at a specific spatial and temporal nexus that is attributed to or affected by a human agent. The term geographic instance is borrowed from studies denoting geographic phenomena, but it is often based on environmental agents found by satellite imagery (Guo et al. 2011; McIntosh and Yuan 2005). This study applies the definition to a geographic-related event as an instance occurring at a geographic location and at a given temporal instance, involving a known human agent.

The geographic instance provides context to develop an anchor for a specific toponym or POI name. Anchor words are special named words stored in a small authority file and used for heuristics analysis (Wacholder et al. 1997). Anchor words for this dissertation are placenames that are matched to a gazetteer and contribute to identifying the state names. Context defines the associations, purposes, and meanings to a text and minimizes semantic ambiguities. Evaluating text that represent POI names and toponyms with no context leaves the text vulnerable to geo/geo and geo/non-geo ambiguities. The encapsulation of a name within a geographic conceptual framework minimizes semantic ambiguities allowing that name and its location to be georeferenced.

A conceptual framework (Figure 9) develops a pragmatic triangulation composed of human agent, spatial, and temporal elements to minimize the common issues that often

degrade the analysis of the GIR process. The intention of this framework is to validate the hypothesis that:

Toponyms and POI names contained within a travel journal are correctly identified and placed in their geographic location based on geographic instances defined by spatial, temporal, and human agent properties.

The hypothesis of this dissertation's conceptual framework reflects Tobler's (1970, p.236; Waters, 2017) first law of geography that defines:

> *"everything is related to everything else, but near things are more related than distant things"*

Tobler's law permits the assumption that POI names and toponyms will geographically relate to other toponyms and POI names in proximity. This presupposition works when a document contains toponyms that are within proximity of one another. A detailed account of an event experienced by the author of the travel journal creates the conceptual framework that anchors an event to a spatial location based on a known temporal and human agent.

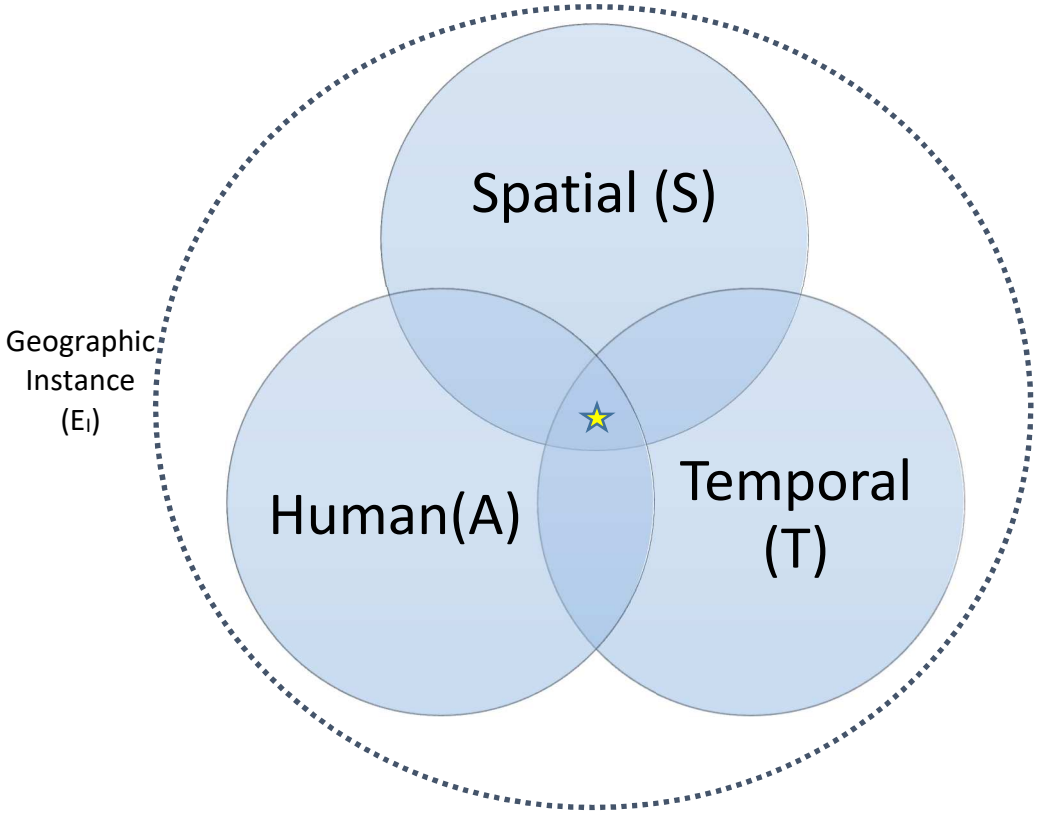

Figure 9 Conceptual Framework. The "star" represent the event's instance (E_I). $E_I \in S \cap T \cap A$

The travel journal used for this dissertation is "The Lost Continent: Travels in Small Town America" (1989) by Bill Bryson. The travel journal incorporates geographic instances of an event based on his actual visits or observations. Figure 10 is an artisitic, cartographic rendering of the route taken by the author along his cross-country drive. The author set out his journey with some predetermination of where he should travel first (towns near Des Moines, Iowa, and near his grandparent's house), but as he ventured further into his travels his destination decisions became focused on his attitude, or the "pull" to that place. The travel journal was selected because it displayed a high number of pull factors that resulted in numerous events set at or in proximity to a setting. It is

also a near national-scale journey conducted in a short period of time, offering a wide range of geographic locations, a variety of toponyms, and a familiar underlying geographical framework. Often times, the author describes an event at a toponym level using towns, cities, or states names, but when the author wants the readers to know the area at an intimate level, he will describe his observations, the use of specific buildings (e.g. resturants, hotels), and his own personal opinions. As the author continues his journey, he provides detailed accounts of his expereinces which vacillated between serious, comedic, and occasionally, offensive. The factors that determine how Bill Bryson decided upon his route and the places visited are influenced by external or internal factors dictating the chosen direction.

Stop *Geographic Instance*

Figure 10 Each stop and route, identified by the stars and dashes, represent a geographic instance

Power of Three

A travel journal contains geographic instances that are defined by the spatial, human, and temporal components. The three components are used because they are known to be inclusive within a travel journal and formulate a symbiotic relationship amongst each other. Meaning, each component within the triad defines the conceptual framework by providing a direct relationship and association with one another. Defining the limits of the three components to formulate a framework encapsulating a geographic instance is especially important when no formal model exists (Achen 2002). In the case of GIR studies on travel journals, very few used models suited to georeference toponyms and POI names. The approach in this dissertation is to select known components that will minimize ambiguities, and identify a text as a place or POI names. The focus on three components prevents the study from becoming either too complicated or oversimplified.

Three components are ideal to maintain a framework that is not overly complicated or simplistic. Figure 11, 12, and 13 are not Venn Diagrams but a simple graphics showing the complexity of the relationships between several inter-dependent factors. Figure 11 displays an oversimplified association of only two variables. With only two variables the A and B has a direct but very simple relationship. This logical conjunction relies on two components to state the validity of a concept; If A=Spatial and B=Temporal are true then the text for a given toponym or POI name is true although it is known that time and space are not the only deciding factors required to disambiguate and georeference toponyms and POI names. Figure 12 displays a more complicated scenario where four components are introduced. When more than three components exist, there is

a direct and indirect relationship associated with all components. The indirect associations are exemplified by C and B which are indirectly related to one another only when A and D are included. The relationships between C and D are not direct and the factors between A and D must be known to provide associations between C and D. Complex relationships are of value, but for a known process, they must rely on a proven model. Figure 13 displays the ideal relationship using three components. By providing three independent variables (spatial, temporal, and human agent) the study accommodates focus, relationship, and efficiency. A, B, and C all must be true for a context to be accepted as representing a geographic text. The use of three components to identify the geographic instance and provide context to a word describing a toponym or POI name maintains a less complicated process. As the process becomes established and proven, additional components are added to build upon this model. This approach also reflects the principles from Occam's Razor, or the Law of Parsimony, which suggests that among competing options or hypotheses, the one with the fewest assumptions should be selected. The interpretation of this principle in the context of this research leads to the model with three components.

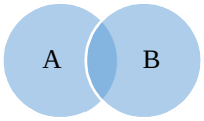

Figure 11 Simple relationship.

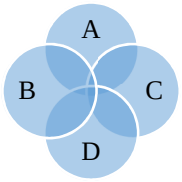

Figure 12 More complicated relationship. Example: C and B do have associations but such associations are dependent on A and D.

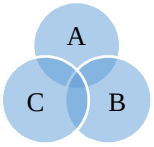

Figure 13 All variables are fully associated with one another.

The dependency between three properties provides a degree of pragmatism to reduce ambiguities and other issues commonly associated with georeferencing POI names and toponyms. The challenge lies in identifying and isolating each event so that a proper evaluation is made. The three properties provide focus, relationship, and efficiency. Focus streamlines the process by emphasizing a limited number of components to georeference a toponym or POI name. The focus on three components provides enough resources to formulate a context to identify a toponym or POI name. Next, each component is related and associated with each other. A direct association is created when three components are used and avoids indirect associations with more than

three components. Finally, with just three components, there are efficiencies in terms of cost, time, and programmatic resources to conduct the study. The overall resources required for the study are minimized, based on the variables most relevant to the study. By focusing on relevant variables, the power of three assists in providing focus, relationship, and efficiency to the project where no or little formal modeling exists, and allows the use of current resources to examine a data source that is unstructured as observed using the travel journal. This approach may have strength in the analysis of text excerpts obtained from the conversations of surveillance suspects, where lengthy blocks of text may not exist and a bag of words approach is suitable

Bag of Words

The unstructured, text-based data source obtained from the travel journal is considered a "bag of words" (Peregrino et al. 2012; Silva et al, 2006). By reviewing data as a bag of words this dissertation ignores grammar and word ordering and places emphasis on georeferencing text representing geographic settings and boundaries. The bag of words concept is not relevant to the travel journal in its entirety, but to each chapter composed within the travel journal. A chapter within a travel journal represents a section that the author deemed important to separate from other chapters. The sections of the travel journal are created at the discretion of the author, but can be related to changes in time, locations, experiences, or themes. Each chapter contains its own bag of words and each bag of words is reviewed separately to identify and georeference all state names, toponyms, and POI names. By separating each chapter as its own bag of words this dissertation can limit the number of texts used for the proximity clustering to provide valid results that are more local or regional in scale rather than having clusters that are state or country in scale. Should the travel journal be reviewed in its entirety the geographic scope would be the entire U.S. and the proximity clustering would be regional at a country level rather than at a state level. The bag of words concept is designed to ignore word grammar and ordering, but the concept can be extended to consolidate sections of a large unstructured document to provide georeference results to the true location of the toponyms or POI names.

Spatial, Temporal, and Human Agent Used for the Dissertation

Spatial, temporal, and a known human agent are the three components required to formulate a context for a text representing toponyms and POI names. The travel journal is a data container created at a specific time frame that contains the toponyms and POI names. External resources will provide context through an entity relationship of the text from the travel journal, the text from a gazetteer, a web locator, Wikipedia, and the author's biography. For the purposes of this study, knowledge acquired for the human agent will aid in providing a presupposition of the author's potential travel originations and location of interests (parent's house, towns with memory associations). The preliminary knowledge of the author's existing places of interest will provide a means to automate a potential origination of the travel journal. Spatial resources are obtained by the NLP application, gazetteers. The NLP process identifies and tags all potential text as a location when parsing the travel journal. The gazetteers and web locator georeference the text identified by the NLP to all possible combination of toponyms and POI names. Temporal resources are obtained from the travel journal and gazetteer. The date the travel journal was published provides a rough estimate of the time frame of the journey. The aid of a human agent to identify the origination of the travel journal, spatial resources to identify and georeference toponyms and POI names, and temporal knowledge to determine the time frame of the travel are used to narrow the list of state names and effectively identify its true location during the GIR process.

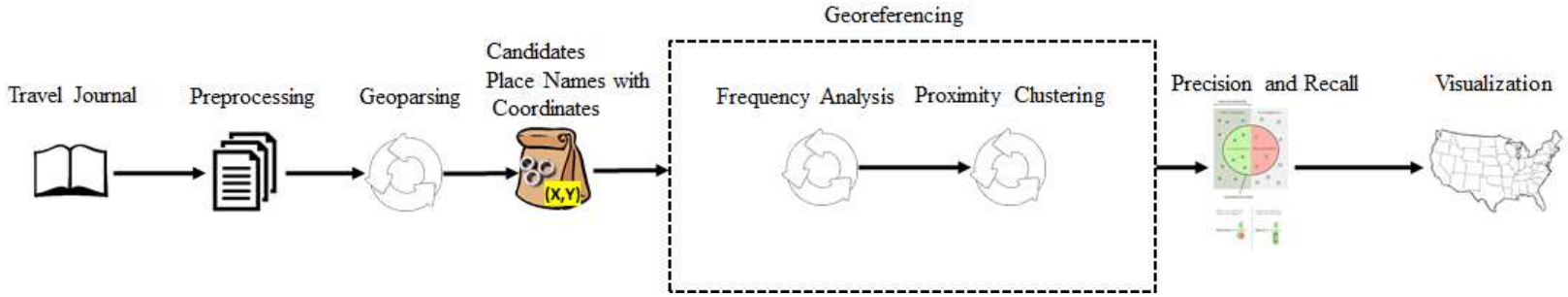

Figure 14 Procedural framework.

GIR Process to Georeference Toponyms and POI Names

This study employs six components to geoparse and georeference a travel journal. The components used by this dissertation are a slight variation from Leidner and Lieberman's (2011; Leidner 2017) reference model. 1) Preprocessing, 2) Geoparsing, 3) Candidate Placenames, 4) Georeferencing, 5) Precision and Recall, and 6) Visualization. Figure 14 outlines and describes the process of geoparsing and georeferencing a travel journal starting with the preprocessing component and ending with the final data deliverable component. The model represents the actual processes and mechanisms to illustrate the importance of the conceptual framework and bring focus to each event located in the travel journal. Table 10 describes the components in more detail. Each component of the GIR process is essential to develop a methodology to identify, collect, georeference, and map the toponyms and POI names contained within a gazetteer. Preprocessing prepares the travel journal for geoparsing and georeferencing tasks; geoparsing identifies and groups all geographic related text; georeferencing adds spatial references to the toponyms and POI names collected; frequency analysis identifies all U.S. state names for a given chapter of the travel journal; spatial proximity clustering analysis groups all toponyms and POI names by proximity; statistical measurements include precision and recall to measure the reliability of the toponyms and POI names collected; and data deliverables provide the maps, visuals and graphs showcasing the analyses and end results. The seven components in this methodology are structured in a linear process to identify and place geographic referenced text from a travel journal onto a map.

Table 10 Description of the eight components of the GIR study

Component	Purpose	Application	Deliverables
Preprocessing	Format data source/travel journal into 29 distinct files allowing the Natural Language Processing (NLP) application to parse and tokenize the text in the document with annotations that defines that text as a location.	Python Script. Travel Journal – data source Geonames is the gazetteer used due to its immediate availability and numerous toponyms. For the United States Geonames has 2.2 million toponyms.	29 Text files for Travel Journal. 51 CSV files listing the toponyms for a state and adjacent states.
Geoparsing	Parsing and tokenization of text from the travel journal. Text with a Named Entity Recognizer (NER) of LOCATION were collected as a candidate toponym.	Stanford Natural Language Processing.	XML files of tokenized elements for each chapter.
Candidate Placenames	A Named Entity Relationship process matching toponyms to a gazetteer.	Geonames – gazetteer for toponyms.	Georeferenced toponyms and POI names were listed in JSON file.
Georeferencing – Statistical Grouping (Frequency Analysis)	Frequency analysis measures the number of state names for a chapter of the travel journal. State names that have the most counts are likely to represent the section	Python - Counts state names based on explicit text found in travel journal and based on Candidate placenames	Output listed in JSON format.
Georeferencing – Spatial Proximity Cluster	Clustering isolates more relevant toponyms into a specific group and add non-relevant toponyms in the remaining groups. Cluster group with the most placenames is more relevant to the travel journal.	K-Means clustering. Offset = 0.5 (range: 0.0 – 1.0) K-Means clustering is an efficient and simple means to isolate more relevant toponyms into a single cluster group.	JSON file.
Reliability Measurements	Precision and Recall.	Precision and Recall measures the success rate of the application	
Data Deliverables	The final output listing the toponyms and POI names.		JSON format
Visual Portrayal	final deliverables - map	Leaflet API.	Web-based map.

Limitations of the Research

The proposed research is a geographic information retrieval (GIR) study that looks at a travel journal as a bag of words and employs a frequency and clustering analysis to validate the correct georeference for a placename. The GIR process for this study emphasizes the proximity, containment, and spatial operations that will place a toponym within an appropriate measure of confidence. The study is concerned with employing a name-based, frequency, and clustering approach to provide context to a geographic text. The goal of this dissertation is to georeference and map all toponyms and POI names contained within a travel journal, but there are several assumptions and caveats that limit the work in this dissertation.

1) True positive toponyms not contained within the primary state and adjacent state names of the travel journal chapter will be considered false positives. An example is when the author is in New York and begins making comparison to toponyms that exist in Iowa. Iowa toponyms will be counted as a false positive for a chapter whose relevant state is New York. The purpose of this dissertation is to obtain toponyms for the geographic area relevant to the section of the travel journal.
2) The primary travel mode is a vehicle, with a linear travel behavior. The analysis does not distinguish mode types within the travel journal used by the author. In most cases the author preferred mode type is an automobile while walking was secondary and used only to traverse short distances. Bus mode was used in one situation when the author traveled to New York city.

3) The author is the primary human agent. The analysis does not attempt to evaluate if the travel mentioned in the travel journal is that by the author or by others. The travel in the journal is described by the author as solitary, and although this assumption may not be wise for certain applications, e.g., geointelligence, group surveillance, it is appropriate for this text. A critical facet of geointelligence group surveillance, are the connections and relationships between many people, each of whom has a specific spatiotemporal footprint". This general issue is reviewed by Medina et al. (2011).
4) All trips occur near the publication date of the travel journal. The year of travel is known to precede the publication year of 1989. The assumption is that the travel took place in fall 1987 and winter 1988, as stated by the author.
5) The order of the places visited will be based on the textual order. This analysis reviews the document as a bag of words, where the places visited in a single chapter are contained in the same bag of words and are presumed to be related, due to the travel mode.
6) The state name with the highest frequency count will be considered the primary geographic state for that travel journal's chapter. The primary state will determine the adjacent states.
7) Abbreviations and alias names will be ignored unless contained in a gazetteer.
8) Toponyms are georeferenced if they are contained within the primary state and any adjacent states. Toponyms are not referenced if the state exists outside the predefined framework or by distance from the primary state defined.

9) State names explicitly mentioned within the travel journal will be used. It is known that not all state names will be adjacent to the primary state of a given chapter. Collecting state names that fall outside the list of states adjacent to the primary state will minimize false negatives.
10) The gazetteer used is current to the time of the study and not based on the time the trips were made by the author.

Limitations in Frequency Analysis

Frequency and clustering analysis are proven mechanisms for GIS research, but there is an inherent risk of distortion when using count-based data. One application related to the Google search engine, PageRank (Page et al. 1999) demonstrates how people manipulate an application whose premises are based on weighted counts of an item searched. PageRank (ibid.) is a popular ranking tool implemented by the search browsers to determine the most likely website related to a search. The methodology ranks the importance of a web page based on the number of backlinks, or other web pages referencing to the primary page. The ranking system reflects the popularity of the link and rates its importance. The more a web page is accessed the higher that web page is ranked. The methodology used by PageRank serves to promote the search engine's purpose, but exposes a vulnerability that allows organizations to manipulate the system and boost searches in their favor. People who want their website to be placed high in Google's search list will often employ a robot to continuously hit their website via the search engine to improve their ranking. Care must be given when using frequency

analysis to avoid unwanted or unnecessary ranking of a toponyms. This is not to say that there is a malicious intent to distort the facts contained within a travel journal, but it is a declaration that distortions occur. Distortions come in the form of geo/geo or geo/non-geo ambiguities that result in the system incorrectly determining a state name because the erroneous toponym candidate was mentioned frequently, due to the author's reminiscence of events at a different time and place. Care must be taken not to rely too heavily on frequency analysis as the primary georeference for this study. The purpose of the frequency analysis is to determine the primary and adjacent state names of a travel journal section, and aid with the use of proximity clustering in the georeferencing of the toponyms and POI name collected from the travel journal.

Conclusion

This dissertation provides a mechanism to develop and implement a conceptual framework to collect sufficient data to minimize ambiguities and vagueness to properly place a toponym and POI name. NLP and NER are the main linguistics mechanisms for this study, based on methodology and results of previous similar studies. This study employs both NLP and NER, but with focus on a specific topic (geography), to parse specific words that relate to a potential placename of a setting, and reference those words to a specific spatial point or area on a map. The intent of this research is to provide a geographic resolution to the problem of identifying toponyms and POI names from a travel journal, to demonstrate the importance of metadata resources such as gazetteers, and to employ a combination of frequency and clustering analysis to correctly place all

toponyms and POI names.

The conceptual framework presented in this chapter describes the encapsulation of a geographic instance and the processes required to improve the toponym resolution. Studies have developed geo-parsing and geo-referencing techniques for political and administrative toponyms, but this study addressed the importance of improving the toponym resolution using frequency and clustering analysis. This research employs existing procedures for toponyms and extends it to include POI names.

METHODOLOGY

Georeferencing the placenames from a travel journal includes six components. The process is linear with the first two components (preprocessing, and geoparsing) focusing on the preparations and placenames categorizations, and the remaining four components focusing on the analytics, measurement, and deliverables. The flow of the process (Figure 14) starts with the travel journal, preparing the travel journal and gazetteer in preprocessing, geoparsing all placenames using the Stanford NLP application, adding coordinates to all candidate placenames, georeferencing the placenames, measuring the reliability of the data using precision and recall, and visualizing the results. Table 11 describes each component and its purpose for this dissertation.

Component	Summary	Purpose
Preprocessing	• Convert travel journal to electronic text. • Download gazetteer from Geonames. • Separate the gazetteer into fifty-one files for each state and district. • Each of the fifty-one file will contain a toponym for that state and the primary state's adjacent states. • Locate and list all previous places of significance related to the author.	• Prepare documents for geoparsing and georeferencing tasks.
Geoparsing	• Execute the Stanford NLP application for each state. • Export output as an XML file. • Save output in its own directory.	• Identify and tokenize all geographic related text from Travel Journal.
Candidate List	• Python File Name: geo_Reference.py. • Origin Locator. • Read XML file from geoparsing where NER= "LOCATION". • Named Entity Recognition based on gazetteer and Nominatim.	• Add spatial references (e.g. coordinates) to toponyms and POI names.
Georeferencing – Frequency Analysis	• Two Analyses. • First, count the number of occurrences for each state name found associated to the toponym candidate. • Second, associate the toponyms with the states explicitly located in the travel journal. • Total the frequency counts from first and second analysis together to get a final count. • The toponyms with the highest count will act as the main toponym for the chapter and determine the primary state and adjacent state for the next chapter.	• Identify the primary toponym for a given chapter to be used as an origination toponyms for the next chapter.
Georeferencing – Proximity Clustering	• K-Means Clustering. • K-value: elbow of a dendogram graph • Cluster with most toponyms will assume to contain the toponyms related to the chapter. • Cluster with equal number of toponyms need to be reviewed. • All other clusters will be omitted from the final deliverable.	• Group all toponyms most likely associated to the chapter of the travel journal based on proximity analysis.
Precision and Recall	• Precision (positive predictive value) • Recall • F-Score	• Measure the reliability of the toponyms collected
Visualization	• List of all toponyms from final cluster in JSON format. • Map of all toponyms with straight line connection of toponyms.	• Visuals and graphs showcasing the end results.

Table 11 Summary of the seven components required to plot the location from a travel journal.

Preprocessing

Preprocessing prepares the document and reference materials for the natural language processing and georeferencing applications. Preprocessing produces and formats electronic documents of the travel journal and gazetteer so that they are readable by the computational analysis. Figure 15 shows the deliverables from the preprocessing stage:

1. Separate the travel journal into 28 distinct electronic text files based on the 28 chapters.

2. Separate the gazetteer into fifty-one files for each state and the District of Columbia (DC). Each of the fifty-one files will contain toponyms for that state and the primary state's adjacent states.

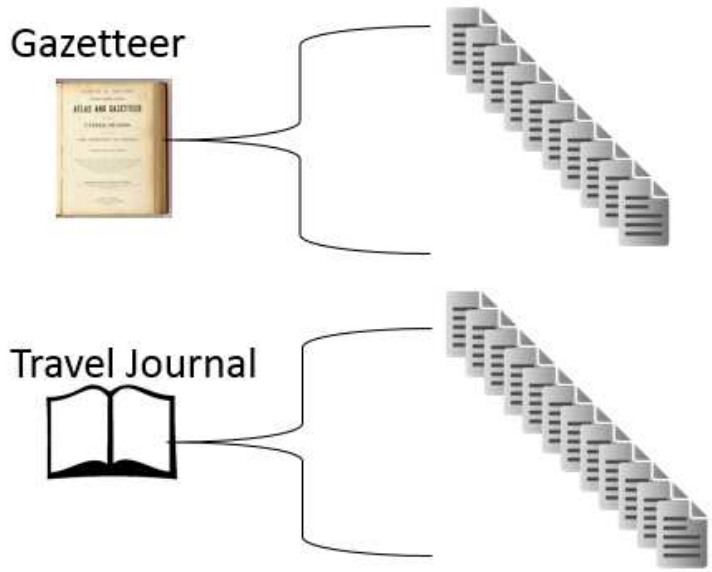

Figure 15 The gazetteer and travel journal are separated into different files based on states and chapters respectively.

The Stanford NLP application cannot parse through long documents using the hardware resources used by this dissertation. To allow the NLP application to geoparse the travel journal larger chapters are split into smaller files (Figure 16). The NLP application on the computer can parse through smaller documents.

The gazetteer is separated by state names and includes data on adjacent states (Figure 17). Reading through one gazetteer file is resource intensive, and by separating the gazetteer into smaller state-specific files the name matching process is focused on a few states rather than all 50 states in addition to DC.

Figure 16 Creating separate files for each chapter of the travel journal

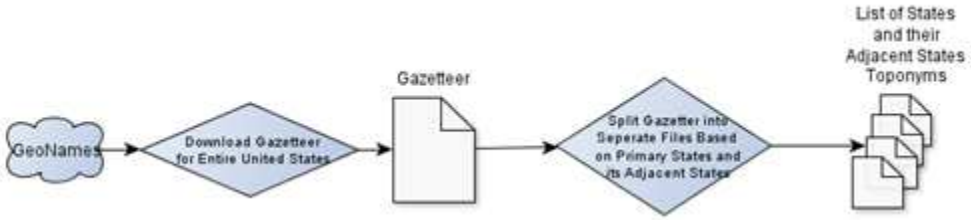

Figure 17 Seperating the gazetters into many files alleviates performance and "out of memory" issues.

Author's Origination

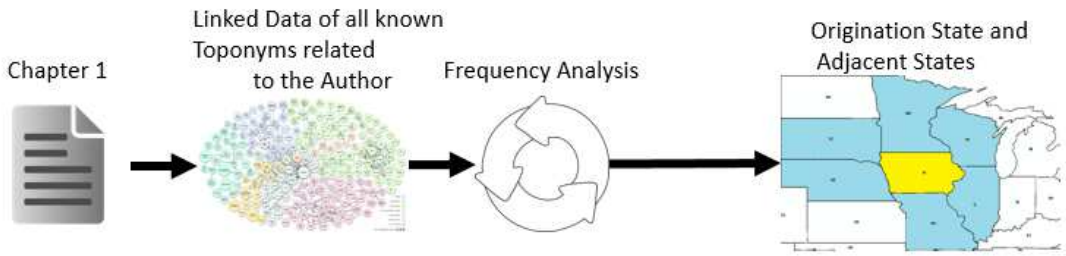

Figure 18 The origination is determined by matching location of importance to the author with that in the first chapter of the travel journal.

Figure 18 outlines the process to identify the origination of the travel journal based on all known locations of interest to the author. Locations of interest are locations specific to the author such as place of birth, residencies, employments, family, or areas of fond memories. The locations of interests by the author are found from Wikipedia and Dbpedia. Using linked data queries (SNORQL and SPARQL) were created to search through Dbpedia and gather all locations specific to the author (Figure 19). A manual search for additional locations was made using Wikipedia. Given the author's popularity in the current time of this dissertation there is sufficient information available about the author. Locations of interests related to people from the general public exists from governmental agencies, and paid knowledge-based sites regarding people's previous history of residencies. All locations found are listed and stored to a JSON formatted file.

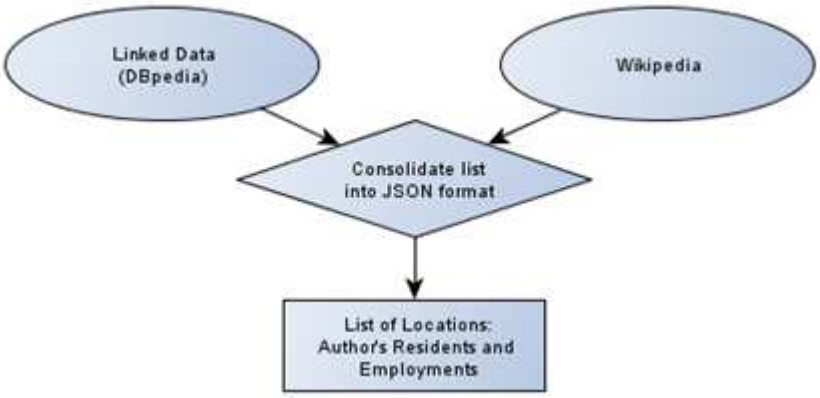

Figure 19 The author's previous places of significance include residencies, employment, and place of birth.

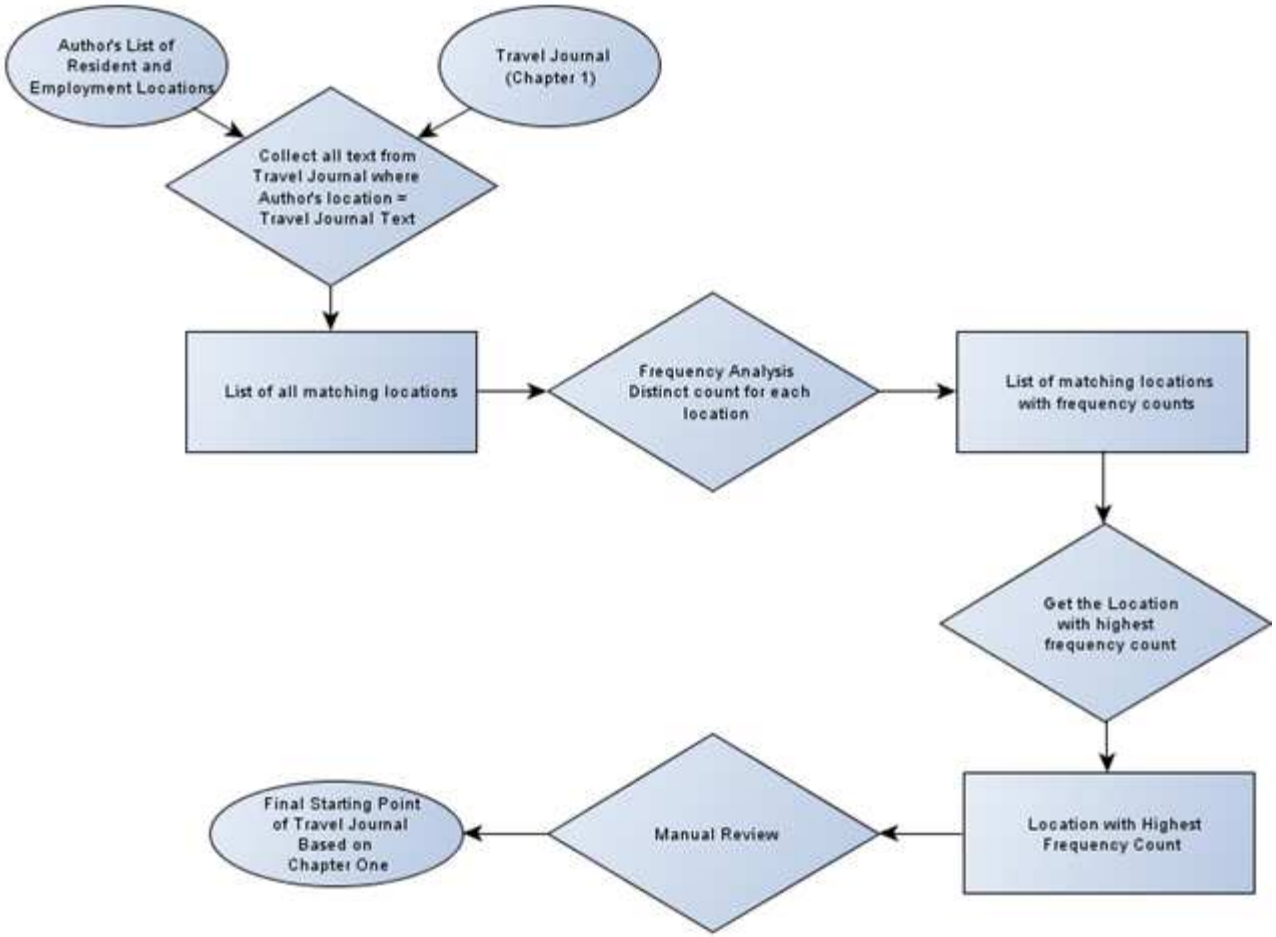

Figure 20 A frequency anlayis was applied to determine if the author's places of significance was the starting point of the travel journal.

The process to identify the origination of the travel journal is based on matching the location of interests' JSON list to the first chapter of the travel journal (Figure 20). The location of interests with significant counts is the origination. A manual review is required to determine if the origination calculated by the analysis is related to the author's previous place of significance. Two toponyms are expected from this analysis: the name of the city or town and the name of the state. The state name narrows the focus of the origination to include the state containing the location of interests and all adjacent state names. Figure 21 displays the goal of using the name matching and frequency analysis to identify the origination of the travel journal.

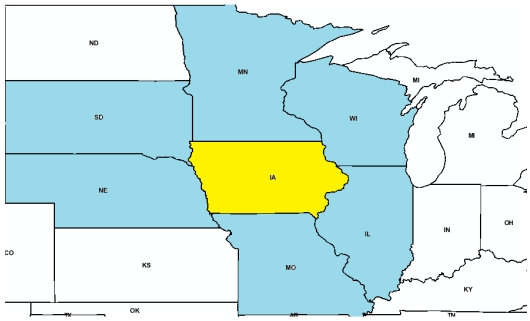

Figure 21 Iowa is the primary state for Chapter one defining the starting point of the travel journal. States adjacent to Iowa are shown in blue.

Geoparsing

Figure 22 Geoparsing text from the travel journal using the Stanford Natural Language Processing application.

Geoparsing identifies and categorizes geographic-related data from text files. The identifications and categorizations of the geographic text is performed by Stanford Core Natural Language Processing[31] (NLP) application (Figure 22). The NLP application reviews all 28 chapters of the travel journal and categorizes all text contained by that document. Figure 23 is an excerpt from the Stanford NLP documentation describing how texts are parsed and tagged as a location using the NLP tool. All text annotated with a NER category of LOCATION by the NLP are reviewed for this dissertation (Figure 24). Tokens encapsulate all annotations and tags to a single text. An annotation is a semantic category or annotation such as named entity recognizer (NER) or part-of-speech (POS) by the NLP application, and identifies a text as a location or a proper noun. The full parameter used to execute the NLP application is in Appendix C. The texts collected are stored in an XML document for the next stage of the process.

[31] http://nlp.stanford.edu/ and http://stanfordnlp.github.io/CoreNLP/

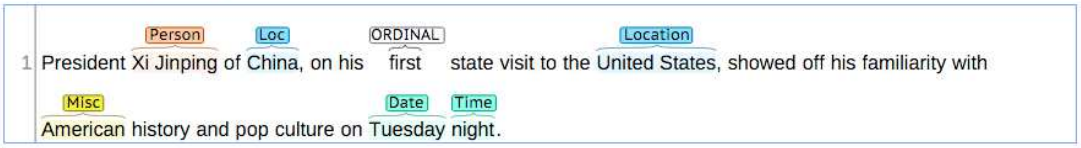

Figure 23 the text of interest is marked with a Named Entity Recognizer (NER) as "Location". Image taken from Stanford NLP webpage: http://stanfordnlp.github.io/CoreNLP

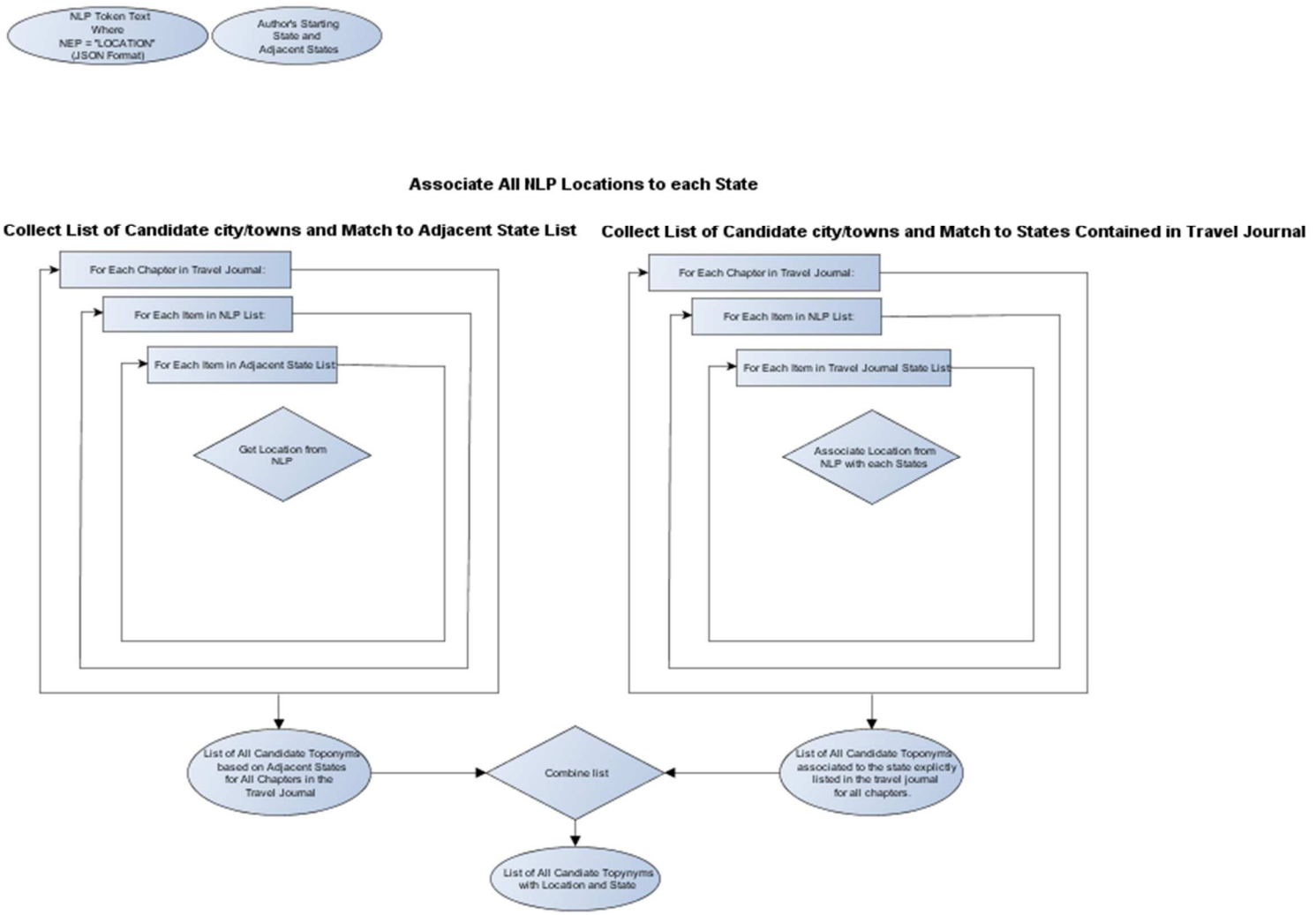

Figure 24 The Natural Language Processing application is required to tokenize the geographic text.

Candidate Placenames

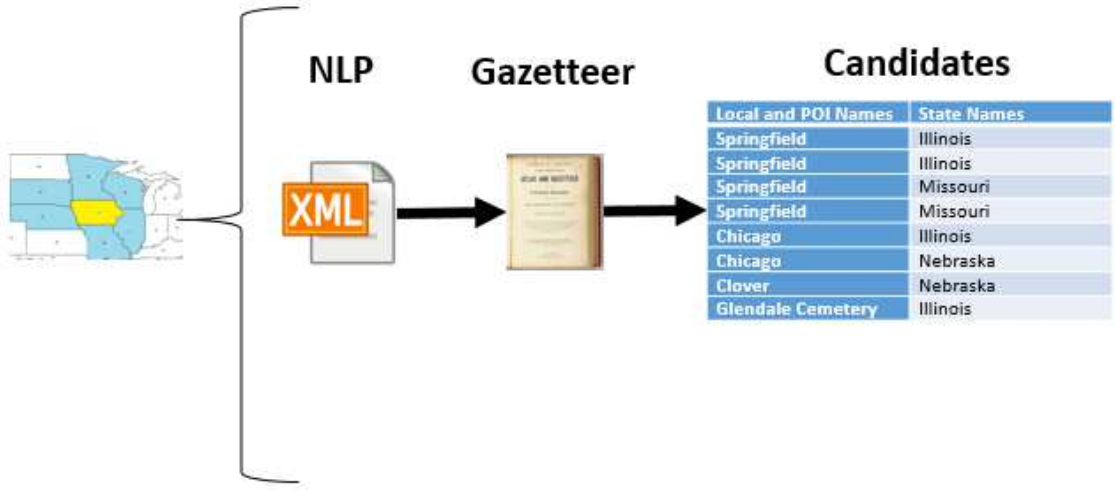

Figure 25 Identify all candidate placenames from a given chapter.

Figure 25 outlines the named matching process to provide a list of candidate placenames. Placenames collected by the NLP application are matched to the name in the gazetteer. All names that are matched are stored into a JSON file that contains the placenames, coordinates, and state names. At this stage of the process no determination has been made of the validity of the placenames. Only that all possible items are matched to a gazetteer, and the coordinate and state attributes for each location are added. The list of all matched items includes:

- Cities, towns, and POI names contained within the travel journal.
- States explicitly contained within the travel journal.
- States indirectly from the cities and towns.

The process of matching the names from the NLP application to the gazetteer is outlined in Figure 26. The first process matches all toponyms and POI names to the

primary state and its adjacent states. A list of all candidate placenames is created based on the existence of the toponym and the POI name for the state names used for the travel journal chapter. If the chapter of the travel journal contains ten toponyms named "Springfield" and three of the states reviewed for that chapter contain the name "Springfield" then 30 records of "Springfield" are generated. The total count of the toponyms and state names collected are saved for the frequency analysis.

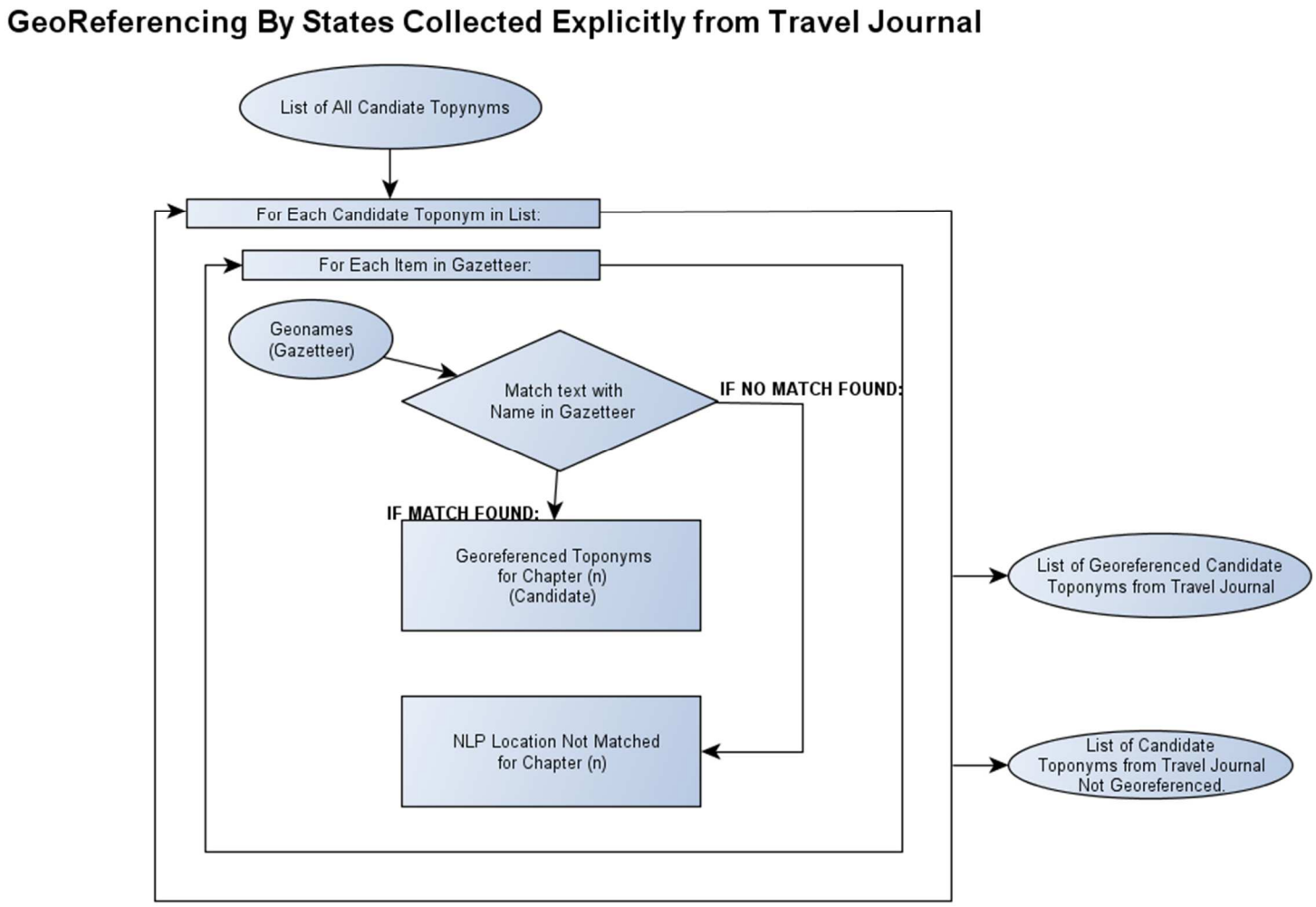

Figure 26 Name match the text from the travel journal entry to the gazetteer.

Frequency Analysis

The frequency analysis is a two-part process that lists all state names for a given chapter of the travel journal. The frequency of a toponym or POI name is fundamental to identify the primary and adjacent state name of the chapter in the travel journal. The state name that has the highest count will act as the origination "anchor" state for the next chapter (Figure 27). The linear travel of the travel journal used allows a presupposition that the starting point in the next chapter is adjacent to one of the states from the previous chapter. The premise of the frequency analysis borrows from Tobler's Law that the author will mention places associated with the current spatio-temporal area of the travel journal more frequently than areas not at or near the author's position.

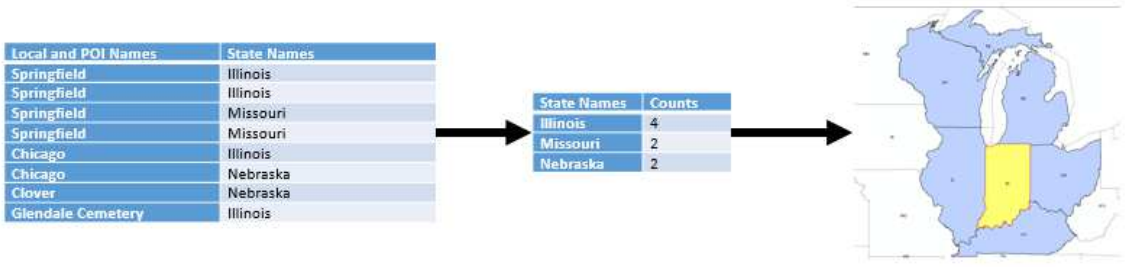

Figure 27 Summary of Frequency Analysis to identify the primary state and adjancent states.

Frequency Part I

The dissertation first addresses all cities, towns, and POI names identified by the NLP application and matched to all possible state names. Table 12 is an example of a toponym candidate that is collected with all possible state names contained. All toponyms and POI names are equally distributed to the state that contain those names.

Table 12 Example of local and POI name georeferenced to a state name.

Local and POI Names	State Names
Springfield	Illinois
Springfield	Illinois
Springfield	Missouri
Springfield	Missouri
Chicago	Illinois
Chicago	Nebraska
Clover	Nebraska
Glendale Cemetery	Illinois

Table 13 and Table 14 groups the frequency count from Table 12. In Table 13 when grouped by placenames and state names no significant counts showed one state as significant. Table 14 removed the placenames and grouped by state names to display a more significant count to a specific state name. At this point, no state name is concluded as the primary state for the chapter. A state may not be adjacent to the list of states used for a chapter. To address states not adjacent to the original list a second frequency analysis is made.

Table 13 Total counts of the toponym.

Local and POI Names	State Names	Counts
Springfield	Illinois	2
Springfield	Missouri	2
Chicago	Illinois	1
Chicago	Nebraska	1
Clover	Nebraska	1
Glendale Cemetery	Illinois	1

Table 14 Total counts aggregated by state.

State Names	Counts
Illinois	4
Missouri	2
Nebraska	2

Frequency Part II

The second part of the frequency analysis reviews the states that were explicitly mentioned in the travel journal. The first frequency analysis took placenames to list all possible states for the chapter of the travel journal, the second frequency analysis obtains state names explicitly named in the travel journal and adds the state names to the list of candidate state names. The names collected from the second frequency analysis process using state names is added to the list shown in Table 15 and Table 16. In Table 15 an additional state, Iowa, was added, this was due to the author's memory of Iowa, and is a state with no relevance to the location described in the chapter. When the state names are grouped (see Table 16) a more significant count is shown to favor one state over others.

Figure 28 shows the intended result based on Table 16, included are the states adjacent to Illinois.

Table 15 The georeferenced toponym candidate is added to the list from the previous analysis.

Local and POI Names	State Names	Frequency Part
Springfield	Illinois	Part 1
Springfield	Illinois	Part 1
Springfield	Missouri	Part 1
Springfield	Missouri	Part 1
Chicago	Illinois	Part 1
Chicago	Iowa	Part 1
Clover	Iowa	Part 1
Glendale Cemetery	Illinois	Part 1
Springfield	Illinois	Part 2
Springfield	Illinois	Part 2
Springfield	Illinois	Part 2
Springfield	Illinois	Part 2
Chicago	Illinois	Part 2
Chicago	Illinois	Part 2
Clover	Mississippi	Part 2

Table 16 Total counts of states after addition of second frequency process.

State Names	Counts
Illinois	10
Missouri	2
Iowa	2
Mississippi	1

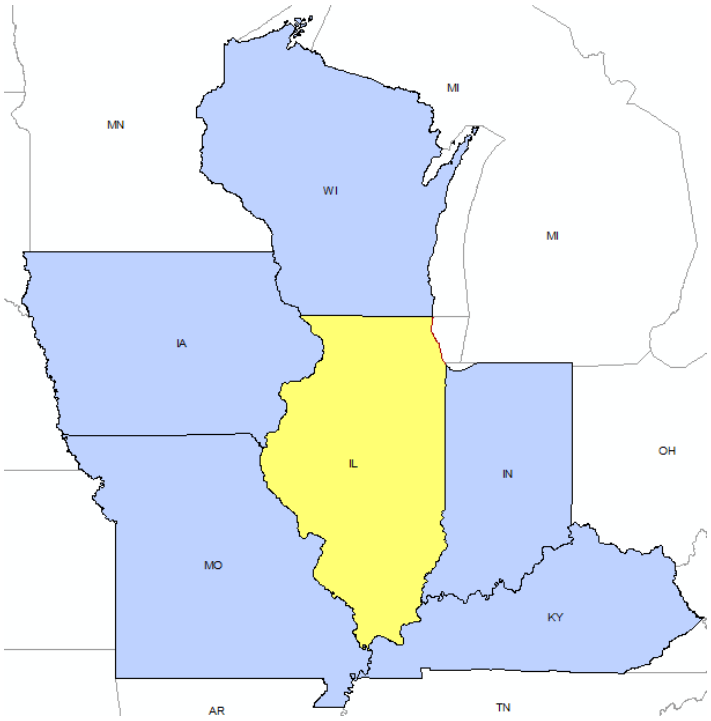

Figure 28 Based on Table 16 Illinois and all adjacent states are used for the next chapter.

Proximity Clustering

The frequency analysis identifies the state names used to narrow the focus of the geographic scope for the toponyms and POI names, but frequency analysis has low confidence in placing toponyms and POI names to its correct location. Semantic ambiguities can skew or inflate the number of invalid placenames by including personal names or locations not relevant to the travel journal. Proximity clustering provides a more confident means to include only placenames relevant to the chapter of the travel journal. Proximity clustering creates a cluster group that contains a significant number of toponyms and POI names that are relevant to each chapter of the travel journal while placenames not relevant to the travel journal are stored into different cluster groups (Figure 29). The proximity clustering analysis used for this dissertation is the K-Means clustering, an unsupervised analysis that contains a pre-defined number of cluster groups (K). A common technique calculating the number of clusters (K-values) is identifying the elbow of a dendogram graph between the coefficient (percent variance) and number of clusters (Ketchen et al. 1996). The average number of toponyms and POI names in the travel journal by chapters is eleven (minimum placename counts by chapters: 2, maximum placename counts by chapters: 39). Because the density of the placenames collected may vary, to avoid having too many or too little cluster groups based on distance this dissertation calculated the K-value by dividing the number of coordinates by two and taking the square root to return an integer. The value three was used as the K-value after taking the average of the number of clusters returned by each chapter. Figure 30 outlines the creation of the empty-sets of a cluster group based on the known K-

values. Other proximity clustering analysis were reviewed such as fuzzy K-Means that cluster each feature to more than one groups with a degree of memberships that measures the feature's relevance to that cluster group; and density-based spatial clustering of applications with noise (DBSCAN) that cluster high density features. K-Means clustering is selected due to its simplicity, density dependency, and single feature to cluster group allocations.

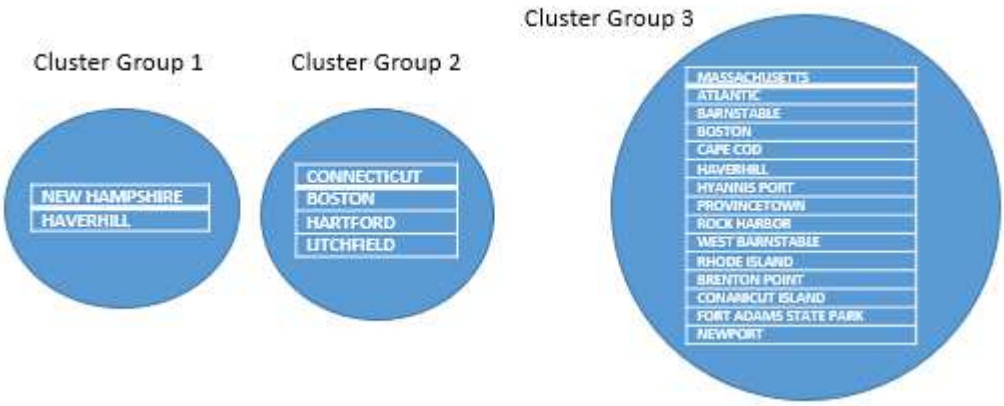

Figure 29 Three cluster groups. Cluster group 3 best represents the travel journal.

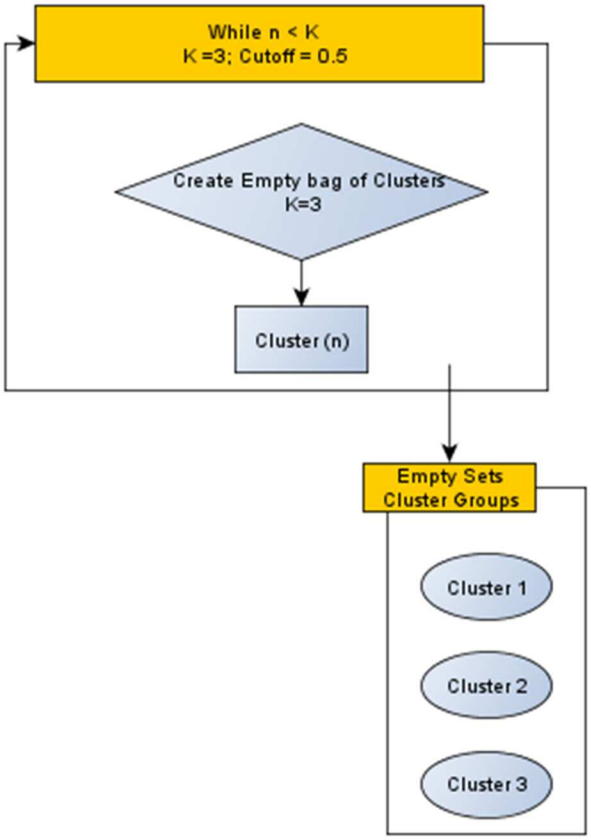

Figure 30 The cluster group is created based on the K-values.

K-Means clustering identifies and groups candidate toponyms based on proximity relationship to the centroid of each cluster group. The word "proximity" is applied to Euclidean distances based on the toponyms placement and their distance to the cluster's centroid. The assumption is that author of travel journals will mention toponyms and POI names that are of proximity to one another in a specific chapter of the journal.

The K-Means clustering process shown in Figure 31 begins with having a known K-value and a centroid is provided to each cluster group. The distance between the centroid and toponyms and POI names are calculated and those placenames are assigned to the nearest cluster group available. The placenames in the cluster groups can shift to a

different cluster group based on the cut-off reassign (Figure 32). Once the clustering process is completed the toponyms and POI names are assigned to the cluster groups, and the cluster groups with the most significant number of placenames represent the chapter of the travel journal (Figure 33). Manual review is needed if the number of items in each cluster group does not display any significant counts. Table 17 displays a cluster example from the study showing toponyms and POI names that are true positives (places visited or observed for the chapter of the travel journal) or false positives. Clustering will not completely remove false positives, but it can minimize the impact of false positives by placing most of them into a separate cluster group.

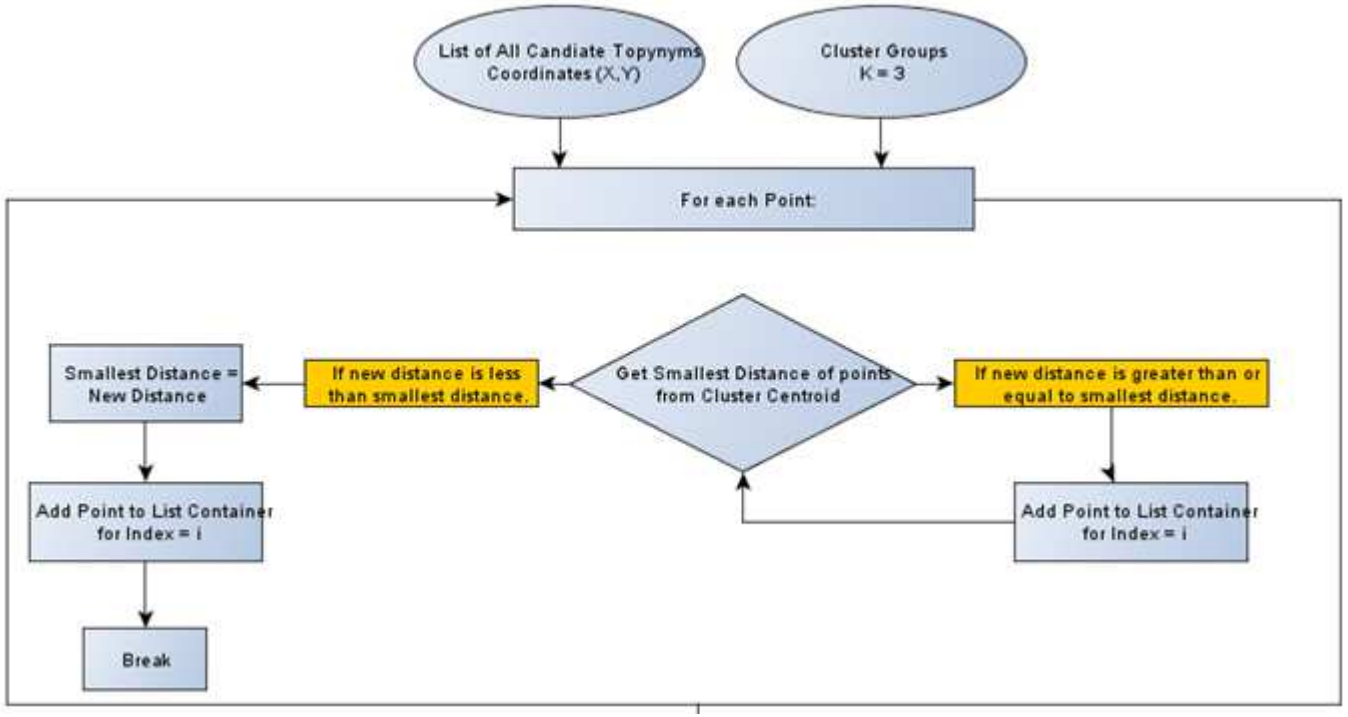

Figure 31 K-Means clustering analyis.

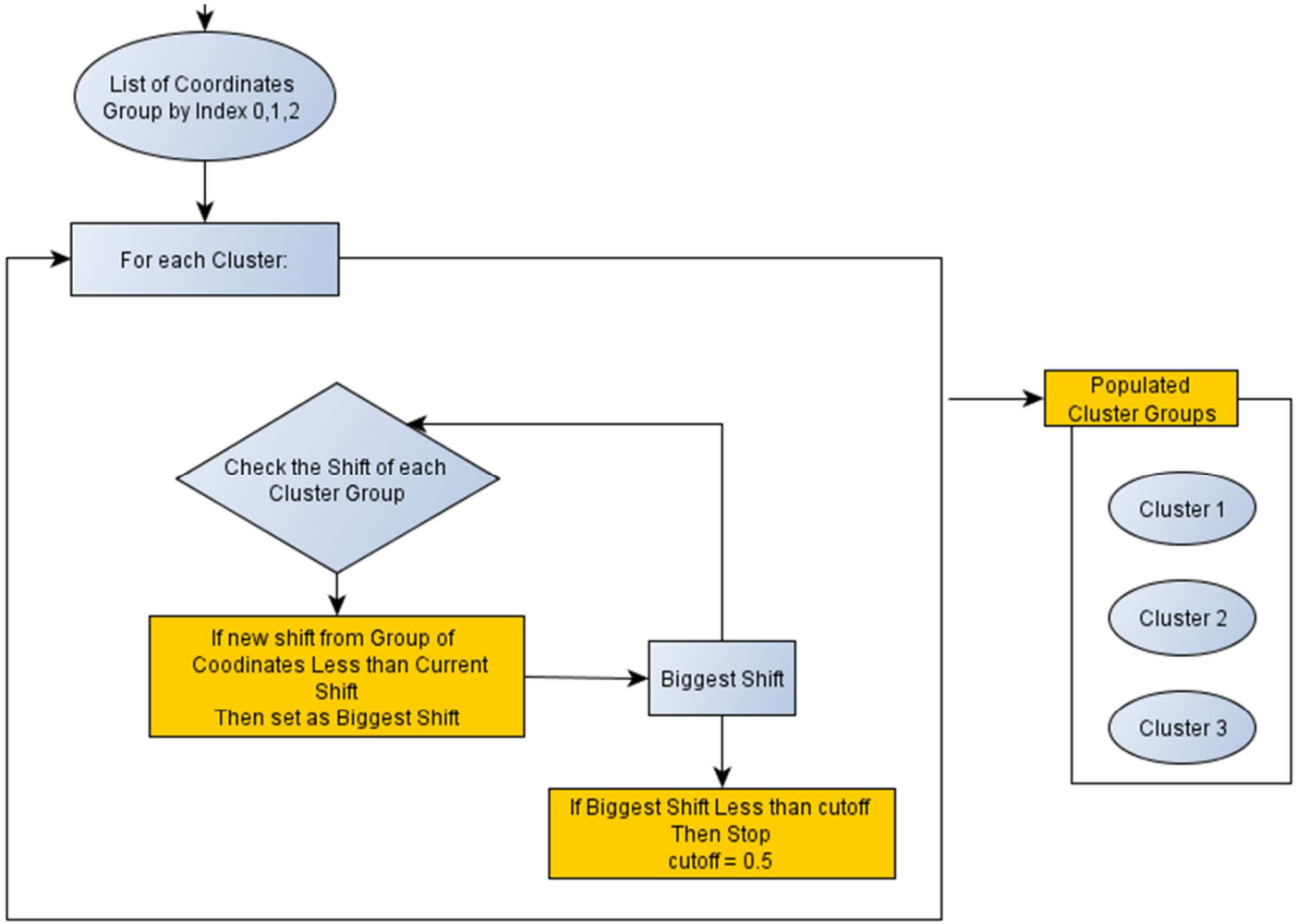

Figure 32 Reviews the distance of each toponym and POI name and associates it to its cluster group.

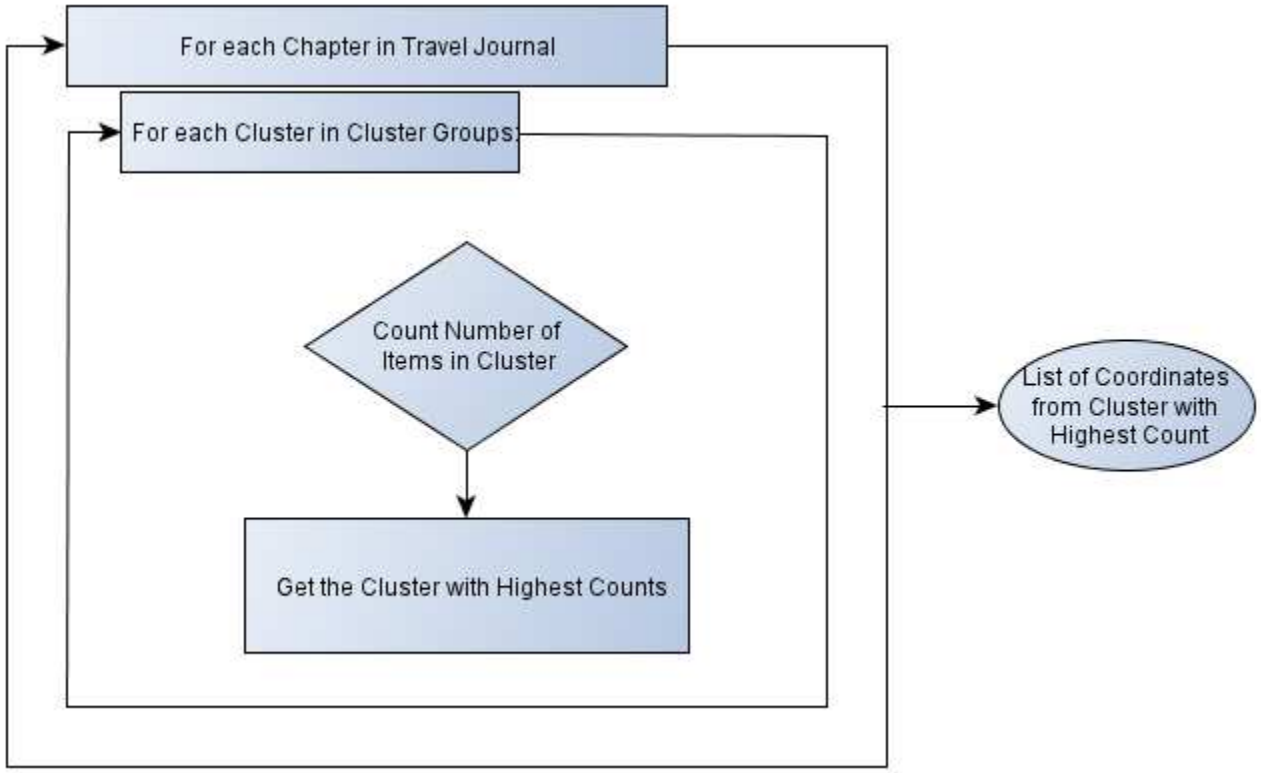

Figure 33 Cluster group with the most toponyms will be considered as representing the travel entry.

Table 17 Example of cluster groups and the number of True and False positive contained within the cluster.

Toponyms from Chapter 15	True Positive	False Positive
Cluster 1		
MASSACHUSETTS	1	
ATLANTIC		1
BARNSTABLE	1	
BOSTON	1	
CAPE COD	1	
HAVERHILL	1	
HYANNIS PORT	1	
PROVINCETOWN	1	
ROCK HARBOR	1	
WEST BARNSTABLE	1	
RHODE ISLAND	4	
BRENTON POINT	1	
CONANICUT ISLAND	1	
FORT ADAMS STATE PARK	1	
NEWPORT	1	
Cluster 2		
CONNECTICUT	1	
BOSTON		1
HARTFORD	1	
LITCHFIELD	1	
Cluster 3		
NEW HAMPSHIRE	1	
HAVERHILL	1	

Four factors determine which one of many cluster groups can represent the chapter:

1) Cluster group containing the most significant numbers of toponym items are used. Table 18 shows “Cluster 2” containing two states with toponyms that best represent the travel entry and will be saved as acceptable toponyms and POI names.

Table 18 Cluster group with significant number of toponyms will be selected as the group representing the chapter.

Cluster 1
ILLINOIS
PARIS
Cluster 2
ILLINOIS
GLENDALE CEMETERY
ROME
IOWA
DAVENPORT
DES MOINES
GLENDALE CEMETERY
MERLE HAY MALL
PARIS
ROME
Cluster 3
INDIANA
AMERICA

2) If a state name in a cluster group that best represents the chapter exists in subsequent cluster groups then the subsequent cluster groups with the state name are added.

In Table 19 cluster 2 is the primary cluster group, but because Cluster 1 contains a state that exists in Cluster 2, Cluster 1 and Cluster 2 will both be used to represent the toponyms of that travel entry.

Table 19 If a state is found in a subsequent cluster group containing a state name that exists in the primary cluster group then those groups will be identified as a valid group for the travel journal.

Cluster 1
IOWA
DES MOINES
OSKALOOSA
PELLA
PRAIRIE CITY
Cluster 2
IOWA
BRIGHTON
BURLINGTON
COLUMBUS JUNCTION
COPPOCK
MOUNT PLEASANT
WAYLAND
WINFIELD
Cluster 3
IOWA
FREMONT
MARTINSBURG

3) Any false positives contained within a valid cluster group will be maintained by that cluster group and identified as a false positive toponym. An example is shown in Table 20 where Paris and Rome are toponyms that exist in France and as its own city-state in Italy. The U.S. Census Bureau[32] lists 20 accounts of Paris and 30 accounts of Rome in the United States and its territories. Names of places can be shared causing confusions to the location of the place. Although there are many instances of the use of the placenames in the US, Paris and Rome were not places visited or observed by the author but were mentioned. Paris and Rome were kept as part of the analysis but flagged as a false positive.

[32] https://www2.census.gov/geo/docs/reference/codes/files/national_places.txt

Table 20 All False positive data will be maintained in the cluster group. In this example, the false positives are "Rome and "Paris" in Cluster 2.

Cluster 1
ILLINOIS
PARIS
Cluster 2
ILLINOIS
GLENDALE CEMETERY
ROME
IOWA
DAVENPORT
DES MOINES
GLENDALE CEMETERY
MERLE HAY MALL
PARIS
ROME
Cluster 3
INDIANA
AMERICA

4) If no significant counts exist among the cluster group a manual review is required. Table 21 shows all three cluster groups significance counts to determine which is the valid cluster group.

Table 21 No significant counts exist among the cluster groups. Manual review is required.

Cluster 1
WASHINGTON
DES MOINES
Cluster 2
WASHINGTON
PACIFIC
Cluster 3
DELAWARE
PHILADELPHIA
WILMINGTON

Precision, Recall, and F-Score

Precision and recall calculates the retrieval rate of all true positive for the toponyms and POI names. Precision shows the ratio of all true positive against the total false positives.

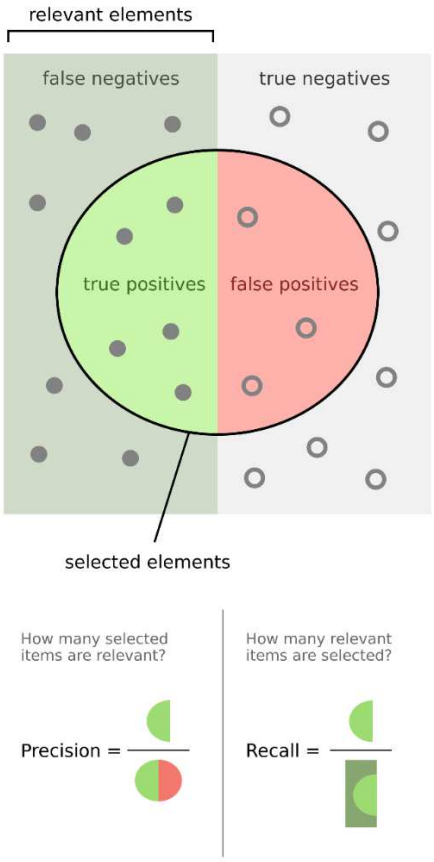

Figure 34 Precision and Recall[33].

[33] https://www.google.com/imgres?imgurl=https://upload.wikimedia.org/wikipedia/commons/thumb/2/26/Precisionrecall.svg/2000px-Precisionrecall.svg.pngandimgrefurl=https://en.wikipedia.org/wiki/Precision_and_recallandh=3636andw=2000andtbnid=eEuqFj-hWs22MM:andtbnh=160andtbnw=87anddocid=rCa1-SuJ1Z_myMandusg=__yVsepWN-eyGYGT51GQlpUtz3D1g=andsa=Xandved=0ahUKEwiB0qulrpHMAhUJ9x4KHRLFAugQ9QEIITAA

The calculation for precision, recall, and the F-Score are:

TP: True positive
TN: True negative
FP: False positive
FN: False negative

Precision = TP / (TP + FP)
Recall = TP/ (TP+FN)
F = 2*(P * R) / (P + R)

Recall shows the ratio of true positives against all false negatives. Precision asks, "of all the toponyms and POI names collected, what is the ratio of those items being true positives". Recall asks, "of all the possible candidates stored in the travel journal, how many were actually identified as true positives". Precision calculates the quality of the existing data, and recall calculates the quality of the data collected. All analysis will be made against the NLP items collected as it was determined to be a more reliable source of reading data from a travel journal than manual searches. Factors influencing precision and recall are "false positives", negative items predicted as positive, and "false negatives", positive items predicted as negative. "True positive" and "true negative" exist as correct results or correct absence of results (Figure 34). Precision and recall provide a separate ratio of all relevant toponyms obtained, but the F-score combines the two ratios to measure the accuracy of the test by comparing results from precision and recall rates.

Conclusion

The reference model for this study is composed of six components. The preprocessing stage identifies the test document and prepares the document for geoparsing. The geoparsing stage retrieves geographic texts from the travel journal. The NLP toolkit focuses on the heuristic and disambiguation of toponyms and POI names. Candidate placenames are matched from a gazetteer and given coordinates for the frequency analysis. At this stage, all toponyms are given coordinate values but no decisions were made as to the actual location of the placenames.

Frequency and clustering analysis are georeferencing tools to identify the location for the toponyms and POI names. The frequency analysis identifies the primary and adjacent states to narrow the geographic scope when name matching toponyms and POI names. The spatial proximity clustering groups potential toponyms based on proximity relationships. Reliability measurements include precision and recall to calculate the application's retrieval rate and measures the quality of the application. Precision measures the true positives against false positives and recall measures the true positives against false negatives. Visualizations provide an assessment as to the quality of the toponym's resolution, and facilitates human interaction to manage any corrections that are needed. Human involvement resolves existing problems and is considered necessary due to the complexity of the process. Human interaction is based on similar efforts used by the intelligence community and private sector to build geographic meaning from narratives. The human involvement and the processing steps assist in the GIR process and improve the resolution of the analysis.

RESULTS

The methodology used in this dissertation incorporates frequency analysis and proximity clustering to identify and group relevant toponyms and POI names contained within a travel journal. The frequency analysis identifies the primary and adjacent states of the chapter, and proximity clustering identifies the toponyms and POI names associated with the chapter. The cluster group that contains a significant number of toponym and POI names represent the geographic setting for a given chapter. The hardware and software used to manage the automation is listed in Appendix B. The evidence associated with the output is presented in this dissertation chapter.

Trip Origination

The origination of the trip is required to provide a starting point for the analysis. The intent of this study is to automate the location of the origination of the trip based on the author's previous and current place of interests (e.g. location of residencies, employments, families). The author's location of interests is based on data obtained from linked data, DbPedia and Wikipedia. Having a set of pre-defined locations can alleviate manual assertions and provide supervised machine learning materials, but the pre-defined locations may not represent the origination transcribed in the travel journal. The results of the initial review returned one toponym ("Des Moines", $k = 16$). In this dissertation, Des Moines is the correct origination, but a manual review was required to verify that the origination from the analysis is correct.

Placement of Toponyms and POI Names

Figure 35 displays the overall goal of this project which is to outline the destinations the author mentioned in the travel journal. The travel described by the author is a nation-wide drive through the United States with many destinations attributed to American smaller towns and cities. During the travel the author described his visits for each town offering notable characteristics personal memories, historical facts, comical comparisons, and personal pride. Figure 36 displays the results from this dissertation of all toponyms with a graduated symbology based on frequency counts, and Figure 37 displays the travel line started and ended at each convex hull generated by toponyms from each chapter. Each point represents the toponym and POI name georeferenced and is comparable to the actual travel in Figure 35.

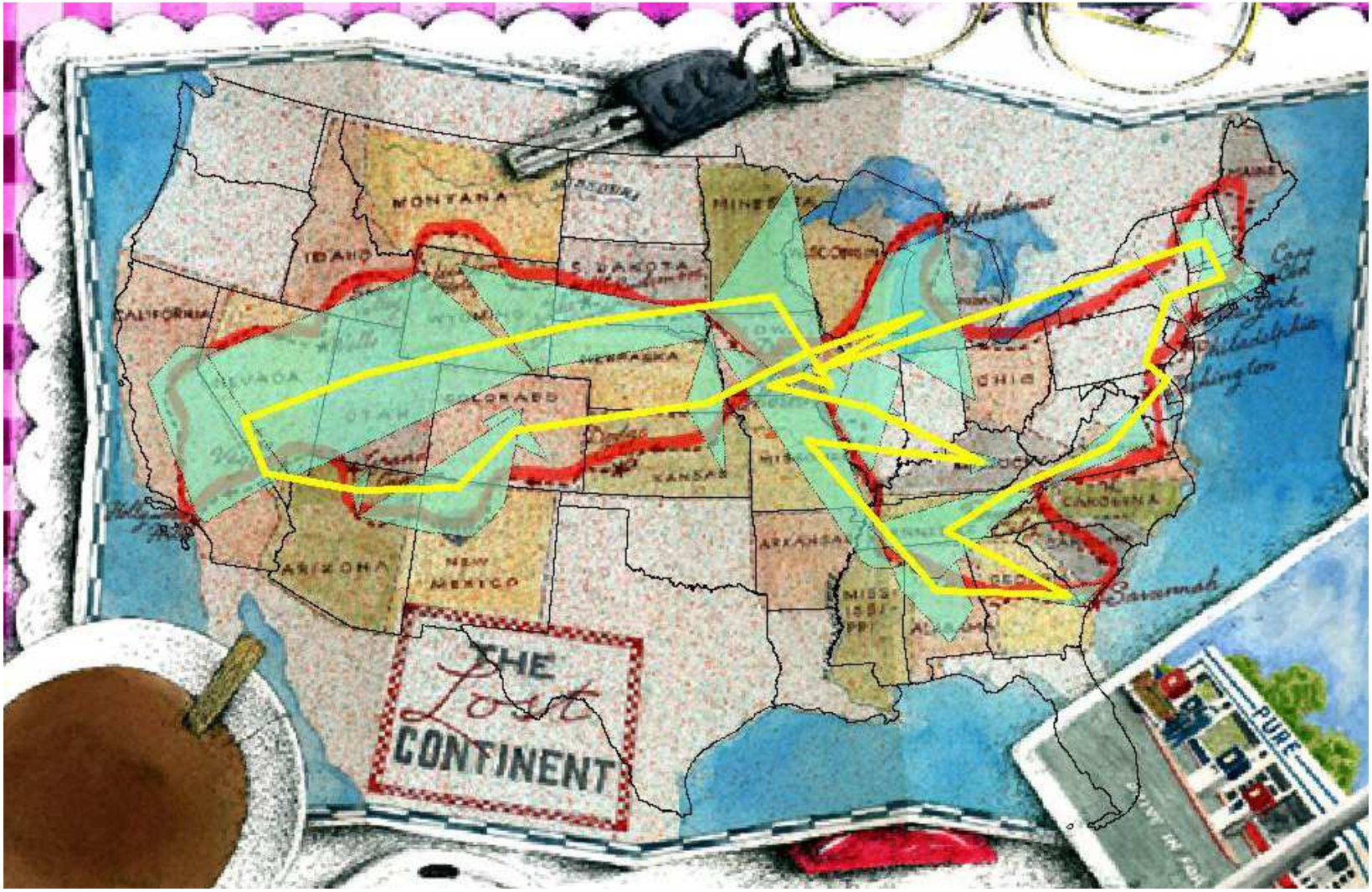

Figure 35 Original outline of the trip taken by the author, Bill Bryson, overlayed with the results of this dissertation.

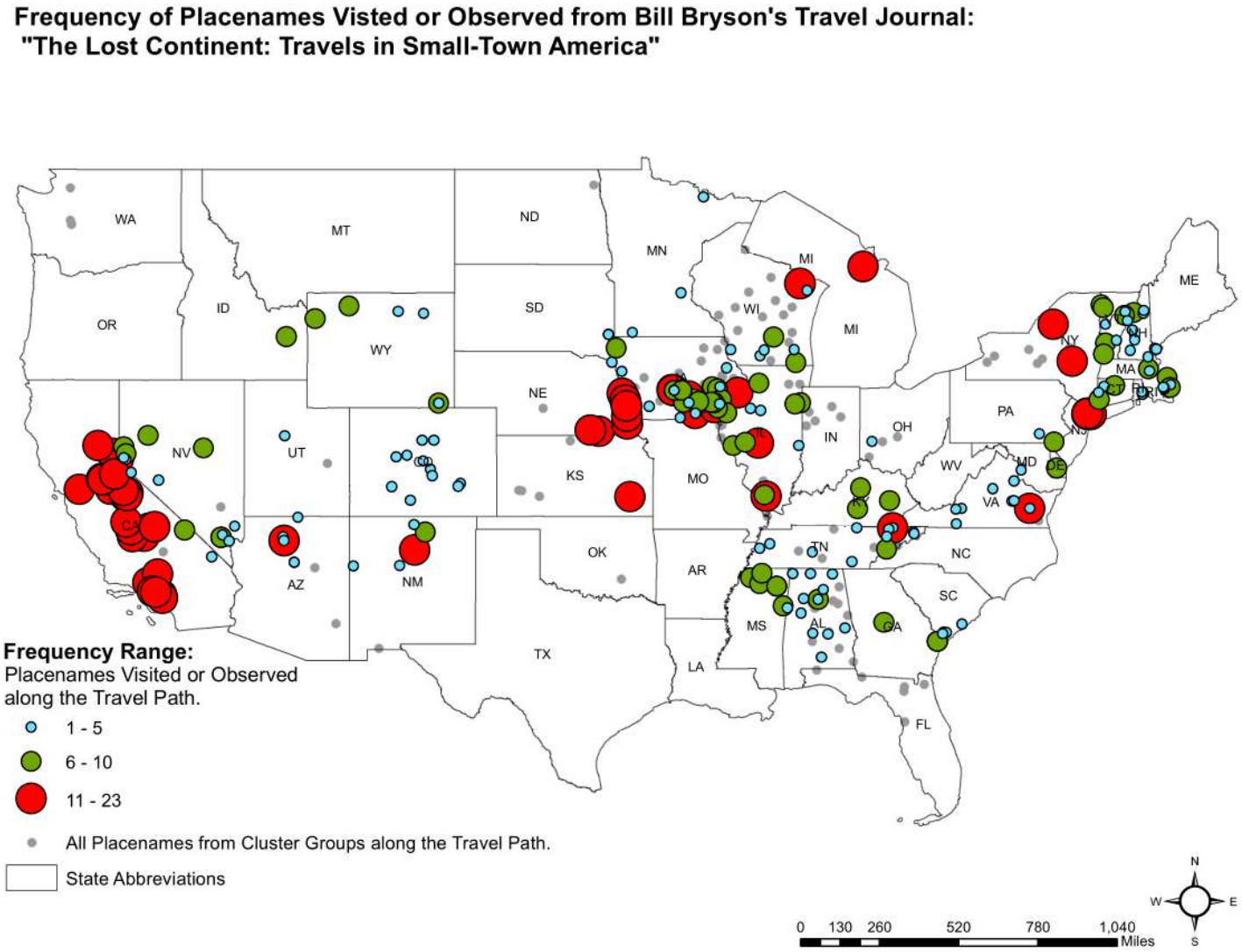

Figure 36 The valid toponyms from the cluster grops are plotted to the map .

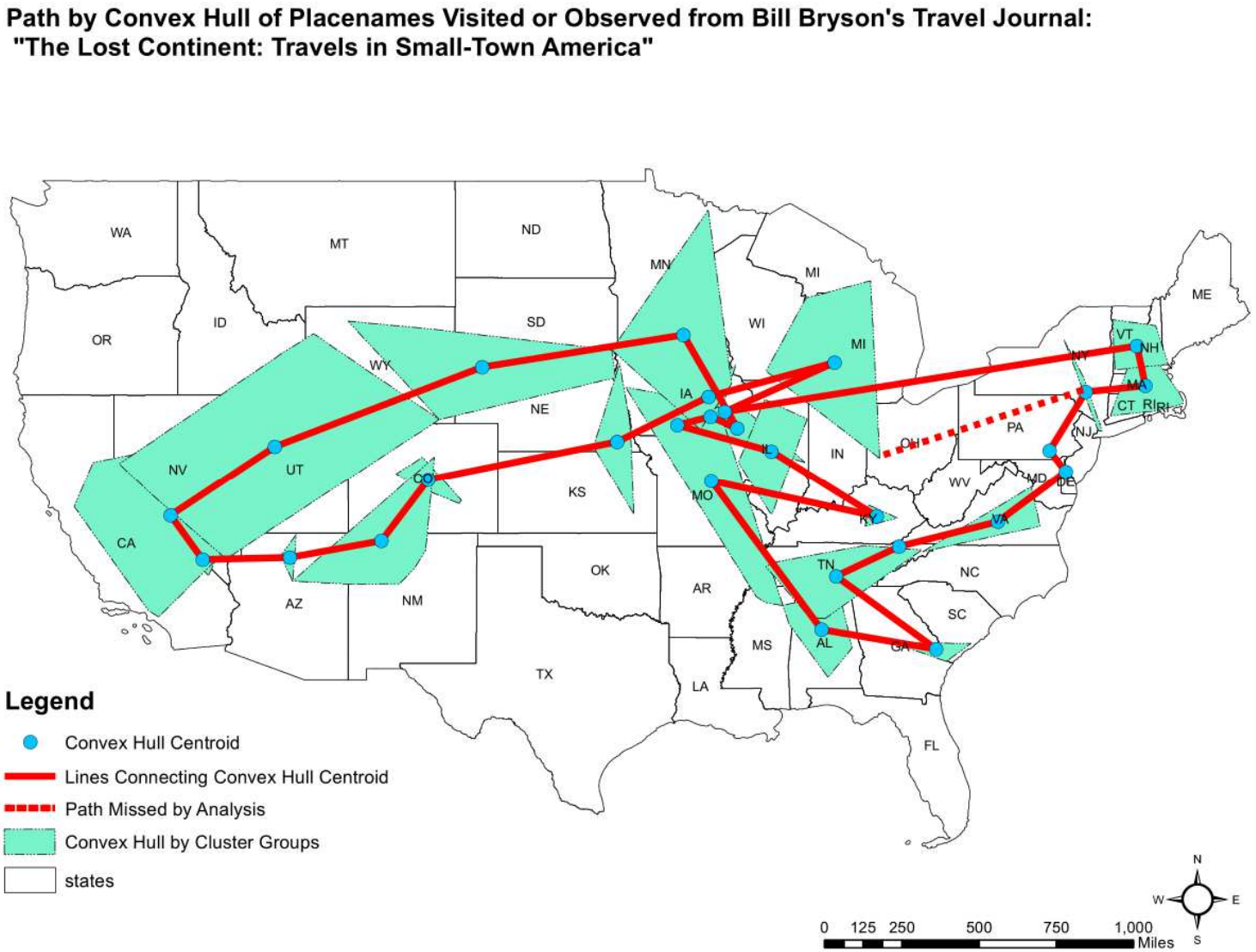

Figure 37 Travel path.

Figure 38 shows the disbursement of all toponyms and POI names (excluding non-U.S. placenames) identified by the NLP Stanford application and grouped by the travel journal chapters. Table 22 breaks down the percentage of states, cities/towns, and POI names in the travel journal. The toponyms in Figure 38 include placenames from flashbacks and comparisons. Each chapter contains at least one toponym and each toponym can belong to the same or a different state since each state can share local toponyms and POI names. Table 23 shows the percentage of toponyms in the travel journal based on the Stanford NLP application. The table shows 1.19 percent (number of toponyms and POI names = 1,287) of the total text in the travel journal (k = 108,142) represents toponyms and POI names. Seven percent (number of foreign toponyms = 91) of the toponyms were foreign names which were not referenced to their foreign countries as only U.S. gazetteers were used to match names. Non-states toponyms (cities, towns, county names) made up nearly 46 percent of all toponyms and state names had 32 percent.

Table 22 Summary of ratio of toponym types from Total toponyms.

States/Total Toponyms	32.25
Localities/Total Toponyms	45.77
POI/Total Toponyms	14.92
Non-US Countries/Total Toponyms	4.97
Non-US Localities/Total Toponyms	2.10

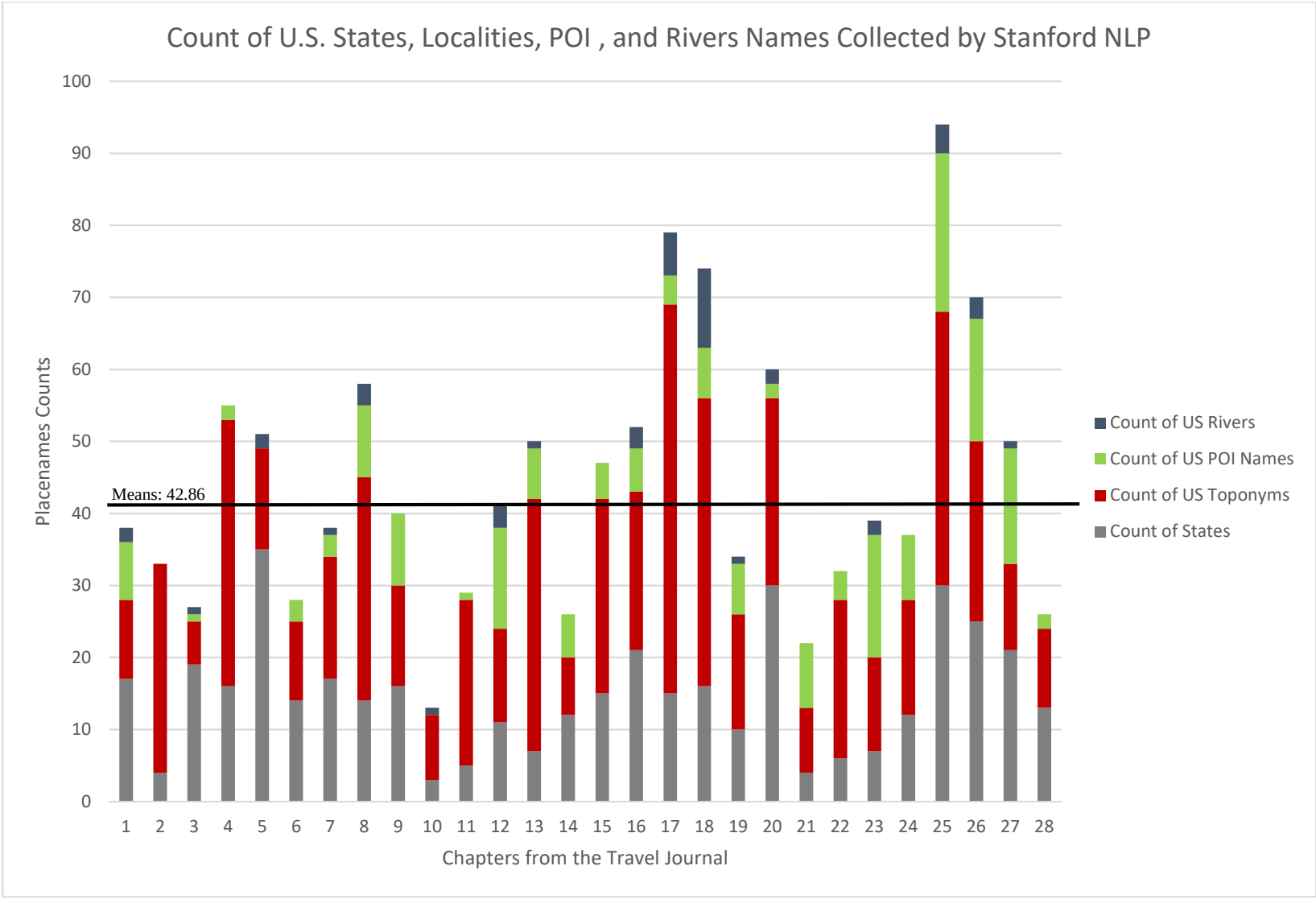

Figure 38 Counts of U.S. toponyms, POI names, and U.S. Rivers by each Chapter.

`

Table 23 Count of Toponyms and POI Names from each chapter in the Travel Journal.

Chapter	Number of Words	Number of Non-Distinct Toponyms	Toponyms/ Word Total	Number of States	Number of Localities	Number of POI	Number of Foreign Countries	Number of Foreign Localities
1	4195	43	1.03	17	11	8	4	3
2	3441	35	1.02	4	29	0	2	0
3	3824	28	0.73	19	6	1	2	0
4	6279	58	0.92	16	37	2	3	0
5	2343	51	2.18	35	14	0	2	0
6	3089	34	1.10	14	11	3	3	3
7	3200	37	1.16	17	17	3	0	0
8	4066	56	1.38	14	31	10	1	0
9	5075	52	1.02	16	14	10	11	1
10	2033	14	0.69	3	9	0	1	1
11	3095	30	0.97	5	23	1	0	1
12	3612	41	1.14	11	13	14	2	1
13	5772	54	0.94	7	35	7	4	1
14	2958	28	0.95	12	8	6	2	0
15	3135	50	1.59	15	27	5	1	2
16	4341	51	1.17	21	22	6	2	0
17	4277	75	1.75	15	54	4	1	1
18	4798	70	1.46	16	40	7	3	4
19	3976	35	0.88	10	16	7	2	0
20	4902	61	1.24	30	26	2	3	0
21	1927	22	1.14	4	9	9	0	0
22	3411	34	1.00	6	22	4	1	1
23	3386	38	1.12	7	13	17	0	1
24	3475	38	1.09	12	16	9	1	0
25	5350	95	1.78	30	38	22	3	2
26	4190	74	1.77	25	25	17	6	1
27	5623	53	0.94	21	12	16	1	3
28	2369	30	1.27	13	11	2	3	1
Total	108142	1287	1.19	415	589	192	64	27

`

Table 24 Distinct counts of toponyms and POI names from the NLP.

Chapter	Distinct Toponym Counts	Distinct POI Names Counts	Distinct Toponyms Georeferenced	Distinct POI Names Georeferenced	Distinct Toponyms not Georeferenced	Distinct POI Names Not Georeferenced	% Toponyms Georeferenced	% POI Names Georeferenced
1	11	5	7	1	4	4	63.64	20.00
2	15	0	14	0	1	0	93.33	N/A
3	10	0	7	0	3	0	70.00	N/A
4	28	0	15	0	13	2	53.57	N/A
5	32	0	25	0	7	1	78.13	N/A
6	12	3	9	0	3	3	75.00	0.00
7	20	3	12	0	8	4	60.00	0.00
8	23	8	10	0	13	10	43.48	0.00
9	23	7	9	0	14	7	39.13	0.00
10	8	0	2	0	6	0	25.00	N/A
11	15	0	13	0	2	0	86.67	N/A
12	17	9	4	1	13	8	23.53	11.11
13	17	7	2	0	15	7	11.76	0.00
14	9	3	6	1	3	2	66.67	33.33
15	21	3	16	0	5	3	76.19	0.00
16	27	4	17	0	10	5	62.96	0.00
17	40	2	8	0	32	2	20.00	0.00
18	36	2	13	0	23	2	36.11	0.00
19	15	5	7	1	8	4	46.67	20.00
20	21	2	16	1	5	1	76.19	50.00
21	11	2	9	0	2	2	81.82	0.00
22	16	2	11	0	5	4	68.75	0.00
23	16	5	9	2	7	3	56.25	40.00
24	16	3	9	2	7	1	56.25	66.67
25	45	13	31	4	14	9	68.89	30.77
26	43	5	14	0	29	5	32.56	0.00
Total	547	93	295	13	252	89	53.93	13.98

`

Precision and Recall

Table 24 displays the total number of distinct toponym and POI name counts for each chapter in the travel journal based on the Stanford NLP application. The distinct counts for each chapter are used to calculate the precision and recall. This dissertation collected and georeferenced nearly 53 percent of the total available toponyms and 4 percent of the total available POI names (Table 25).

Table 25 Summary of placenames count visited or observed.

	Resources	Total Counts	Relevant Counts	Placenames Preserved from Travel Journal (%)
Toponyms	Travel Journal	883	585	
	NLP	813	489	84
	Dissertation	311	311	53
POI Names	Travel Journal	232	232	
	NLP	73	64	28
	Dissertation	10	10	4

The precision and recall in Table 26 shows a high precision but low recall with a harmonic mean (F-Score) of 52 percent. POI names are removed from the precision and recall seen in Table 27. The precision remains unchanged but the recall increase to 45 percent and F-score increased to 60 percent. The 88 percent precision is like previous peer-reviewed studies (Table 28) and shows that the present methodology can place

`

toponyms and POI names obtained from the geoparsing application. This dissertation can improve the peer-reviewed applications by automating the georeferencing process, identifying the high-level toponyms such as state names that contain subsequent toponyms, and identifying toponyms and POI names across a large county like the United States.

The 30 percent recall is related to the toponyms and POI names missed or not matched to a name in the gazetteer. Nanba et al (2009) downplayed the importance of recall for their study deciding that precision is more important. This study shows that recall is very important as it portrays the success of reviewing the available toponyms contained within a corpus. If too many viable candidates become false positives it can distort an analysis or partially georeference a travel entry contained within a journal due to insufficient data. The success of improving recall rates relies heavily on gazetteers and web locator management to be more inclusive of POI names.

`

Table 26 Four percent of the POI names were georeferenced which significantly increased the false negatives and lowered the recall and F-score.

Precision, Recall, and F-Score (Toponyms and POI Names)	
Reliability Measurements	Results
Precision	0.88
Recall	0.30
F-Score	0.52

Table 27 POI Names are omitted from the analysis showing the recall and F-Score increasing.

Precision, Recall, and F-Score (Toponyms Only)	
Reliability Measurements	Results
Precision	0.88
Recall	0.45
F-Score	0.60

Table 28 Reliability Measurements from previous GIR studies.

Application or Processes Name	Statistical Accuracy Outcome	Source
Nominator	92% (precision)	Wacholder et al. 1997
Web-A-Where	80% - 91% (precision rate)	Amitay et al. 2004
Word frequencies	80% (accuracy)	Verma et al. 2011
Supervised machine learning	73 - 85% (accuracy)	Hu, Ge
Gazetteer Classification	78.5% (predictive accuracy training	Garbin, Mani 2005
Supervised machine learning using GeoCLEF for Travel Blogs	86.7% (precision), 38.1% (recall)	Nanba et al. 2009
Microtext Geoparser	.90 (F-Score) .99 (precision, toponyms) .94 (recall, toponyms)	Gelernter and Balaji, 2013

`

Toponym and POI names counts are calculated from the travel journal, NLP Stanford application, and this dissertation. Figure 39 shows all toponym counts from the three sources. The gap between the toponyms collected by the dissertation and toponyms stored in the travel journal is noticeable compared to the gap between the NLP and travel journal which aligned close to each other. Figure 40 focuses on toponyms but only those relevant to the travel journal. The gap between the travel journal and this dissertation is smaller. Also, note no toponym count for this dissertation for Chapter 17. This dissertation failed to capture Pennsylvania as a state for the chapter and the result causes the frequency analysis to travel from New York to Iowa. Next, Figure 41 displays the POI name counts for all POI names that exist in the travel journal. Figure 42 displays the POI name counts that are relevant to the travel journal. In Figure 41 and Figure 42 the POI names captured by this dissertation was not significant. The POI name counts between the travel journal and NLP showed a significant difference, but the NLP can recognize POI names as a location or organization. In this dissertation, the POI names did not exist in the gazetteer used or if it did exist in the gazetteer the POI name in the travel journal was spelled differently from that in the gazetteer. This dissertation collects sufficient number of toponyms to outline the path of the travel journal. Including the POI names would improve the path, but an authoritative reference source for POI name is required.

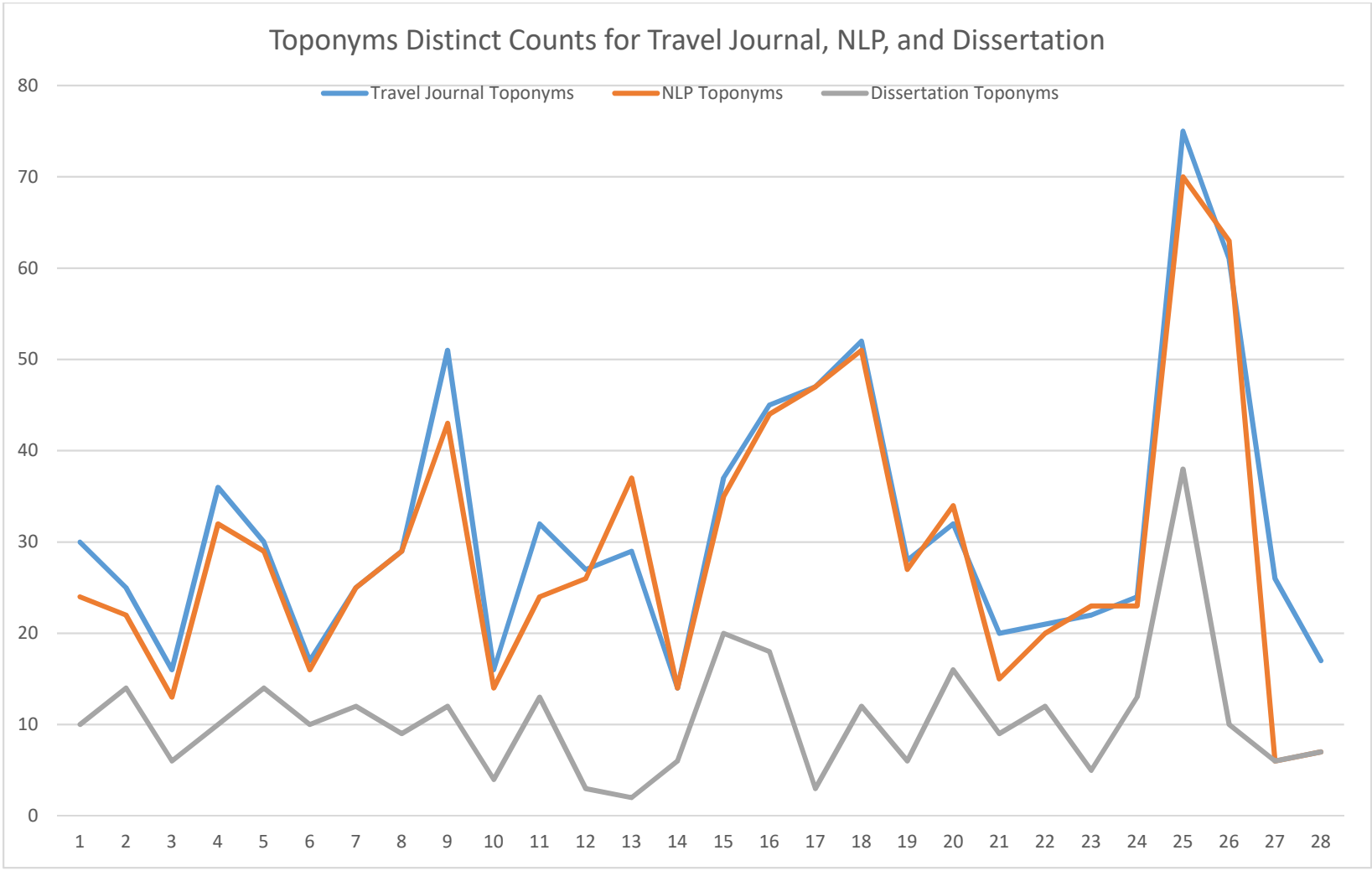

Figure 39 Toponyms Distinct Counts for each chapter of the Travel Journal, NLP, and Dissertation.

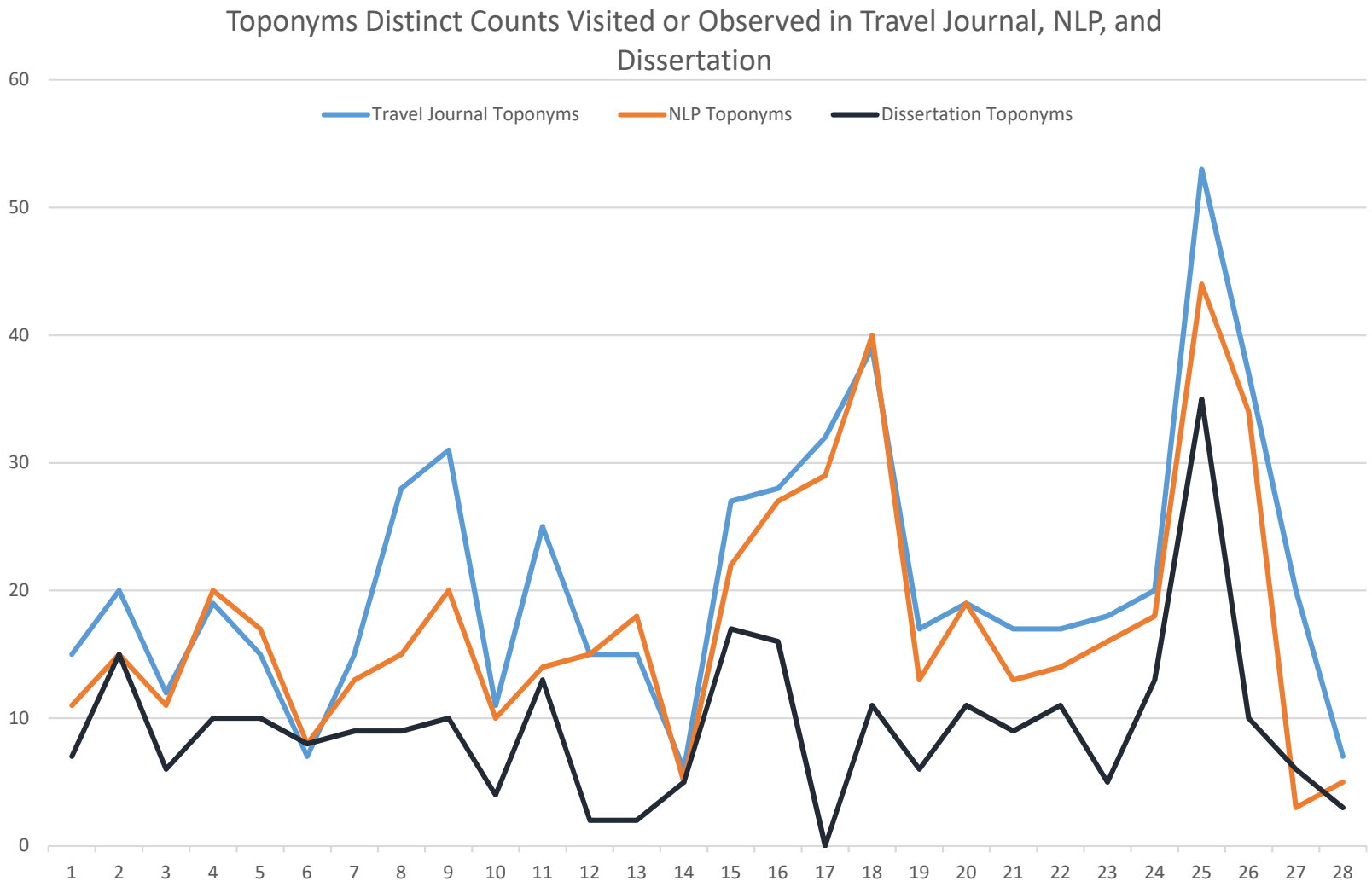

Figure 40 Toponyms visited or observed for each chapter.

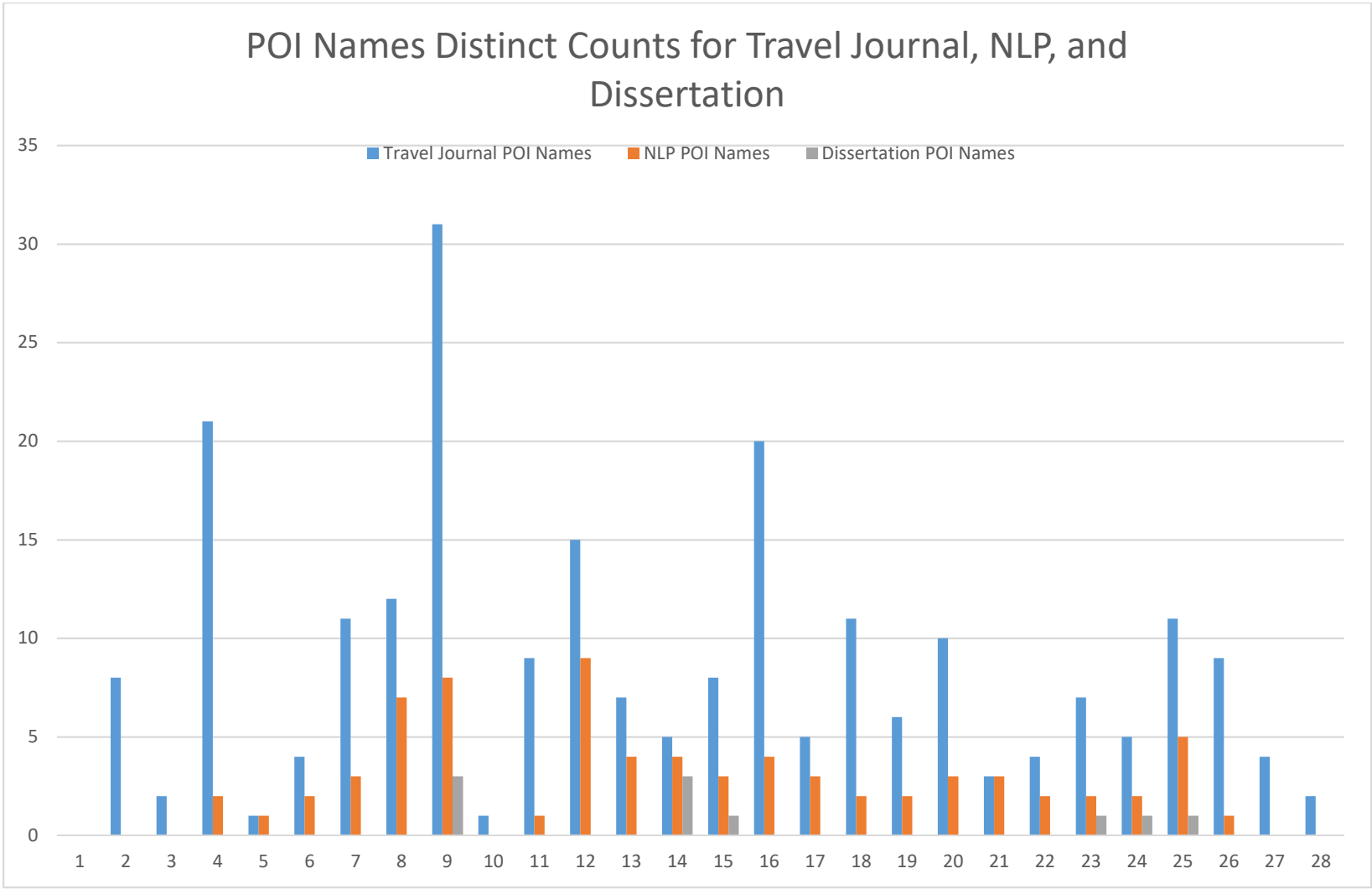

Figure 41 POI Names Distinct Counts for each chapter of the Travel Journal, NLP, and Dissertation

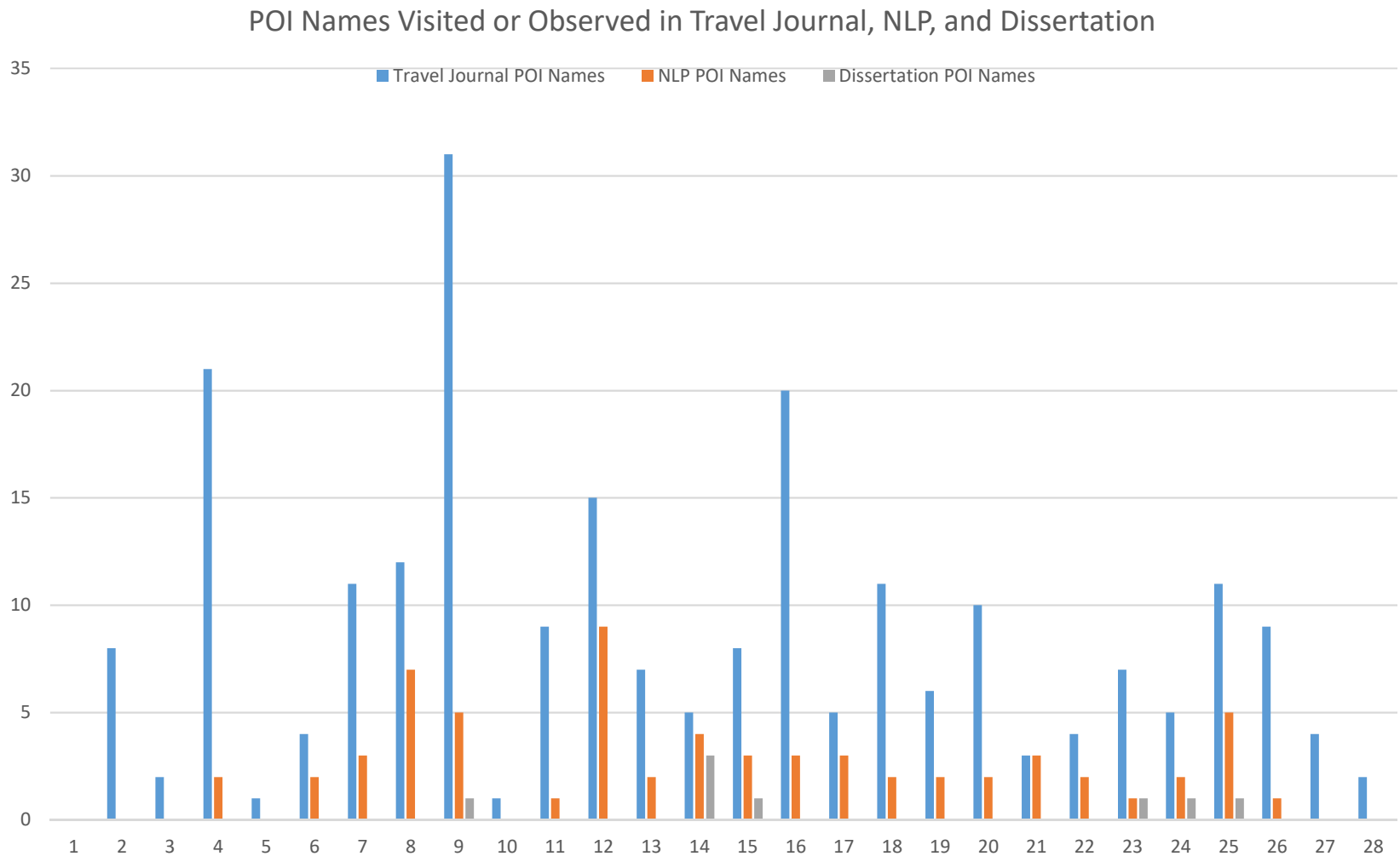

Figure 42 POI Names visited or observed for each chapter.

`

Issues and Caveats

The results have shown precision comparable to peer-reviewed studies, but the recall rate is low. The low recall was anticipated based on previous studies issues with semantic ambiguities, but some issues were unexpected (for example, the NLP application was unable to parse the last two chapters). This dissertation found five issues that negatively impacted the results:

1) Inability to name match all toponym candidates.
2) Inability to name match all POI name candidates.
3) Semantic ambiguities.
4) Flashbacks and Comparisons.

The inability to name match all toponym candidates contained within the travel existed when names in the travel journal did not match the name in a gazetteer due to slight differences in spelling or omissions. The gazetter used to name match toponyms is GeoNames which supports many feature types (including hotel names) (Smart et al. 2010). Valid names such as Alexandria, Virginia, were omitted from the study because the name did not match the gazeteer due to additional text used by that gazeteer (e.g. “City of Alexandria”). A check was added to the program to match a toponym with “Town of” or “City of” if the original name contained no match to the gazetteer, but future study is required to list all possible naming convention of a toponym. Table 29 contains a list of toponyms and POI names included in the study (True Positives) and those omitted (False Negatives).

`

Table 29 Example of toponyms and POI names included and omitted (with reasons) from the analysis. "X" means included or omitted.

Toponyms Chapter 12	Included	Omitted	Reasons for Omissions
Alexandria		X	Format of name differs from gazetteer.
Annapolis		X	Name in gazetteer is different than what was used in book. "City of Annapolis".
Baltimore		X	Format of name does not match name used in Travel Journal.
Brooklyn Bridge	X		
Capitol building	X		
Capitol Hill		X	Name does not exist in gazetteer.
Chesapeake Bay		X	Water bodies not found in gazetteer.
Chestertown		X	Name in gazetteer is different than what was used in book. "Town of Chestertown".
Delaware	X		
Des Moines	X		
Griffith Stadium	X		
Indianapolis		X	Indiana is not adjacent to Virginia (the primary state) nor is Indiana explicitly named in the chapter of the travel journal which would trigger a read for all toponyms in Indiana.
Jefferson Drive	X		
Jefferson Memorial	X		
Lincoln memorial	X		
Maryland		X	No city, town, or POI names associated to state mainly due to named recognition issues related to Annapolis and Baltimore.
Monticello	X		
Pacific		X	Name does not exist in gazetteer.
Philadelphia	X		
Potomac		X	Name does not exist in gazetteer.
St. Louis		X	Name does not exist in gazetteer.
Texas	X	X	No city, town, or POI names associated to state due to named recognition issues related to San Antonio.
Washington Monument	X		
White House	X		
Wilmington	X		

`

POI names also had naming variation issues with gazetters, but the issues stemmed from two causes: the temporal disassoications between the time the travel journal was published (year: 1989) and the time this study took place (year: 2016), and the omissions of the POI names in the gazetteer. POI names were matched from the gazetteer, but only an insignificant number of POI names were placed (POI names count = 13).

Semantic ambiguities were anticipated based on peer-reviewed studies related to minimizing geo/geo and geo/non-geo names. Geo/non-geo semantic ambiguities allows personal names such as "Washington" to be tagged as a toponym when it represents a historical U.S. president. Personal names and toponyms can be interchangeable when no context is provided. In this dissertation Hannibal, Missouri is parsed by the NLP application as a personal name rather than a toponym. Identifying Hannibal as a personal name prevented Hannibal from being included in the frequency analysis and caused the state of Missouri to be missed by this dissertation.

Geo/geo semantic ambiguities allow a toponym to be placed incorrectly when the names are identical. Washington, D.C. and the State of Washington were misinterpreted by this dissertation at first by placing Washington, D.C. as a state name. Clustering analysis resolved the geo/geo ambiguities by grouping toponyms that represent the chapter of the travel journal into one cluster group. Figure 43 illustrates three cluster groups with the highlighted toponyms representing geo/geo ambiguities. Cluster 3 correctly represents the chapter of the travel journal by having more numbers of

`

toponyms in the group. False positives exist in Cluster 3 (e.g. Atlantic) and Cluster 2 (Boston) where the names represent a different location.

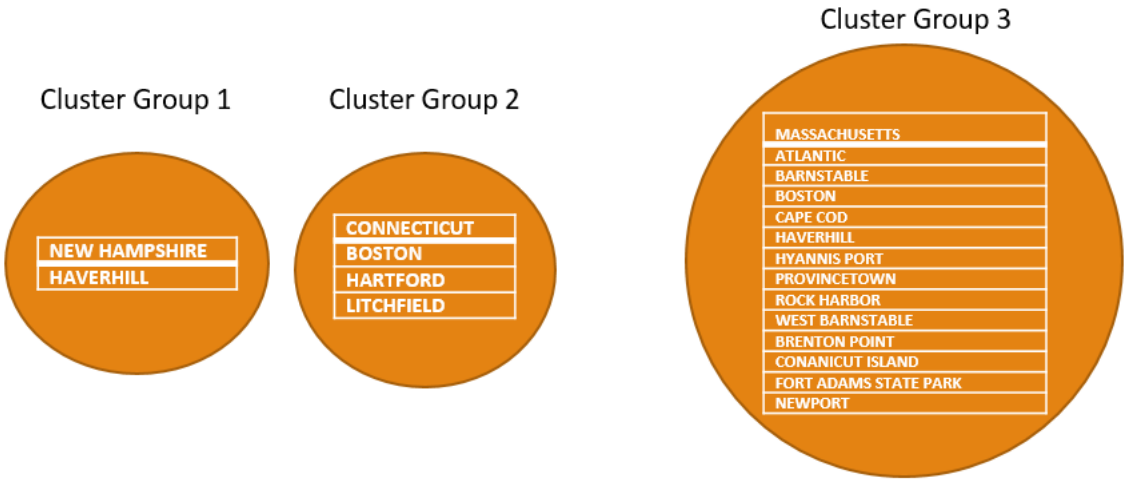

Figure 43 Cluster groups. Massachusetts (Cluster 3) contains significant numbers of toponyms

Last, the author's use of flashbacks and toponym comparison allow incorrect toponyms to represent the travel journal. The author continually has flashbacks of families, past incidents, remembrances, and historical backgrounds when visiting many settings. The author also made comparison of toponyms with his hometown of "Des Moines" or "Iowa" or with foreign countries such as London, Cairo, and Versailles where the same name exists. The use of descriptive analogies and reminiscence are human phenomena that can impact and distort both the frequency analysis and proximity clustering. An example of the author's comparative analogy is in chapter 16 and 17. This dissertation determined that the primary states of chapter 16 contained Maine, New Hampshire and Vermont while the primary state of chapter 17 contained Iowa. Iowa

`

became the primary state when in fact the primary states of chapter 17 are New York, Pennsylvania, and Ohio. Two reasons caused the analysis to determine that the author drove from New England straight to Iowa and not New York. The first, was this dissertation's inability to locate names due to naming variations between the travel journal and gazetteer (e.g. Cleveland, Catskills, Toledo). The second, is the lack of subsequent toponyms related to Pennsylvania. A subsequent toponym (e.g. city, town, or POI name) must be obtained to add a state to the frequency analysis. This rule was added to prevent addition of states that were explicitly mentioned but not necessarily associated with the primary state of that chapter (e.g. Iowa, the author's home state).

`

CONCLUSION

This dissertation extends previous peer-reviewed studies to georeference toponyms and POI names from a travel journal using frequency and proximity clustering analyses. Figure 44 shows the value of using Bryson's travel journal as it has a very broad national coverage for the entire 48 mainland sates. The travel journal emphasizes places the author lived or worked, the author's home town, his travels across the US, and descriptions of his trips. The travel journal allowed an opportunity to develop a mechanism as shown by this dissertation to georeference the toponyms and POI names. This dissertation has achieved measurable success in placing toponyms and POI names from the travel journals, and has done so through a system that is largely automated, and therefore a significant potential resource for conducting similar GIR and georeferencing tasks with large text archives, where speed and efficiency are important

Travel journals are used for this dissertation because they record significant facts by the respondent about a trip made over a fixed period of time. The hypothesis as presented at the beginning of this dissertation presents a vision that recognizes the importance of spatial, temporal, and human agent properties derived from a travel journal to denote who participated in the travel, when the travel occurred, and the locations where the travel took place. The three properties develop a conceptual framework that describes the geographic instances. This dissertation recognizes the events contained within a travel journal to present an opportunity to identify the toponym and POI name as placenames visited or observed by the author of the travel journal.

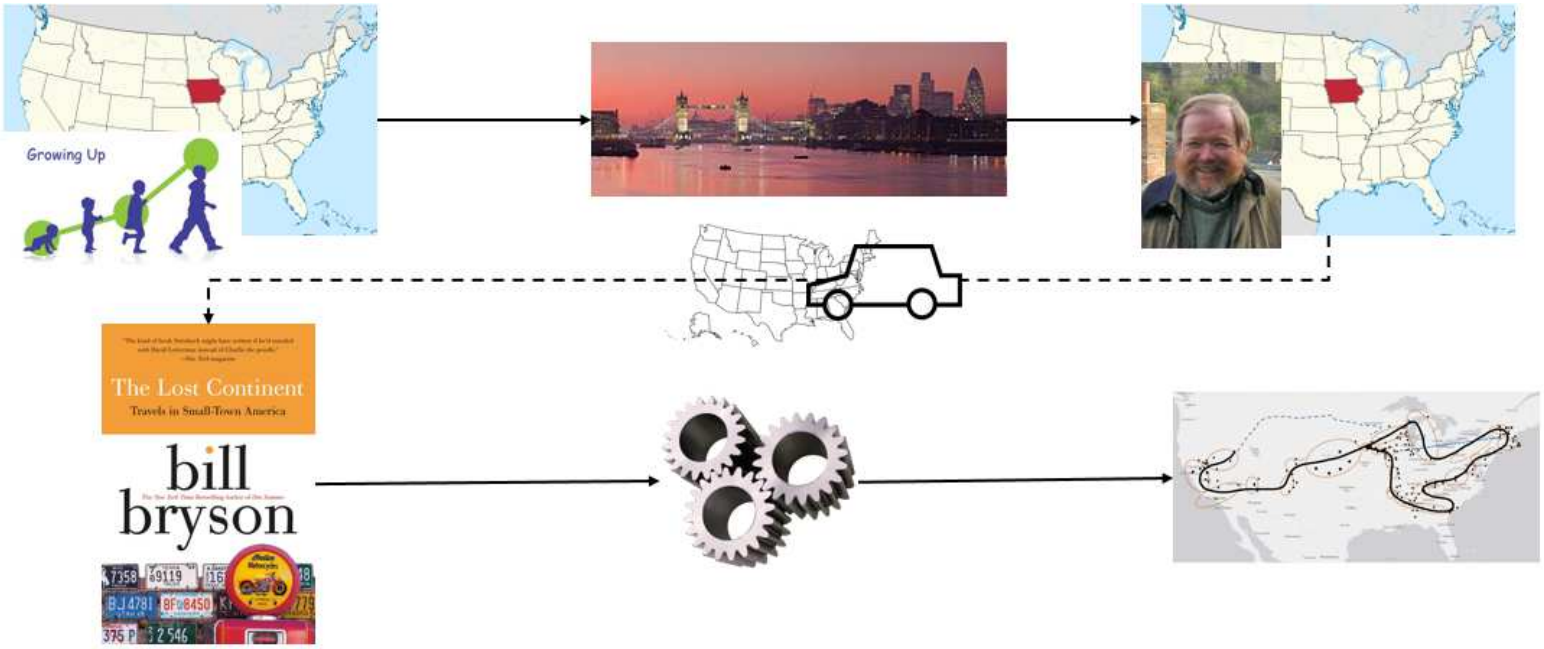

Figure 44 This dissertation means to georeference the author's travel journal.

`

The geographic instance is described as the focal point where spatial, temporal, and human agent merges. The three properties provide the framework to georeference and place the toponyms and POI names related to the travel journal. For this dissertation, the human agent is the author of the travel journal, the spatial components are the toponyms and POI names used to define the location visited or observed, and the temporal element defines the resources required based on the time the trip was taken. These three properties reflect other important work in geography, e.g., Sack (1997, Figure 3) which have a similar emphasis on a nexus of human/socio-relational/temporal factors to provide a sense of geographic place.

This dissertation treats the travel journal as a bag of words ignoring grammar and word order and focuses on georeferencing and placing toponyms and placenames. Toponyms and POI names are identified and categorized by a natural language processing (NLP) application and name matched by an NER process. The Stanford natural language processing toolkit geoparses and groups location text, due to its public availability and wide adoption by other peer-reviewed studies. The NLP is tasked with identifying all text within the travel journal that represents a toponym or POI name. The NER process matches and provides geographic attributes to the toponyms and POI names using a gazetteer. The result is a list of all possible combinations of georeferenced toponyms and POI names that are associated with one or more U.S. state names. The list of candidate toponyms and POI names is used for the next phase involving frequency analysis and proximity clustering. The frequency analysis identifies the likelihood of a state name that contains or partially contains the explicit toponyms and POI names and

`

group toponyms and POI names. The clustering analysis groups all toponyms and POI names by proximity due to Tobler's First Law of Geography by recognizing that toponyms and POI names that are within a certain proximity will relate more than distant placenames. The frequency and clustering analyses minimize negative impacts caused by ambiguities and provide a grouping of proximal geographic candidates specific to the travel journal.

Frequency analysis and proximity clustering identifies state names that contain the toponyms and POI names and group the toponyms that represent the chapter of the travel journal. The state names encapsulate the toponyms and POI names to provide focus and narrow the geographic scope of the analysis. The frequency analysis counts the number of state names containing the toponyms and POI names. The state name representing the chapter of the travel journal is based on the candidate list of the toponyms and POI names. By grouping the list of candidates by state names, a significant count determined the primary state and all adjacent states for the next chapter. Frequency analysis identifies the state names but does not identify or place toponyms and POI names.

`

Proximity clustering groups relevant toponyms and POI names based on the distance to the cluster's centroid. Proximity clustering analysis builds upon Tobler's First Law of Geography which states that "everything is related to everything else, but near things are more related than distant things", and groups all toponyms and POI names based on proximity. A primary cluster group is the group that contains a significant number of toponyms and POI names. More than one cluster group can be used if the state names in the subsequent cluster groups exist in the primary cluster group. This concept is applicable when the mode type for a travel journal is highly linear. The advantage of employing a combined approach of frequency and clustering analyses is that it provides a naïve approach to group toponyms and POI names based on the frequency of a state name and the proximity of the placenames. The disadvantage of this combined approach is that the toponyms and POI names must exist in both the travel journal and gazetteer to provide data matching and data to georeference.

The methodology for this dissertation incorporated six components of the GIR process: 1) Preprocessing, 2) Geoparsing, 3) Candidate Placenames, 4) Georeferencing, 5) Precision and Recall, and 6) Visualization. Each component of the GIR process is essential for the methodology to identify, collect, georeference, and map the toponyms and POI names contained within a gazetteer. Preprocessing prepares the travel journal for geoparsing and name matching; geoparsing identifies and groups all geographic text; candidate placenames create a list of potential placenames and add spatial references; frequency analysis identifies the U.S. state name for a given chapter of the travel journal; spatial proximity clustering groups all toponyms and POI names; precision and recall are

`

reliability measurements; and visualizations provide the maps and graphs of the results. The components in the methodology represent a linear process to identify and place geographic referenced text from a travel journal onto a map.

This dissertation has an 88 percent precision rate and a 30 percent recall rate. The recall rate measures the application success in georeferencing all toponyms and POI names that are contained within the travel journal. The recall rate is lower than the precision rate due to the NLP not tagging POI names as a location and the name matching between the POI names and the gazetteer failed to produce results. The precision rate measures the application success in georeferencing all toponyms and POI names that are collected and stored in a bag of words. The precision rate shows that this dissertation can place the toponyms and POI names when the geoparser and name matching produce significant results. When the spatial, temporal, and human agent properties exist, this dissertation yields a precision rate comparable to other peer-reviewed studies.

This dissertation delivers upon its intent to identify POI names within its spatial, temporal and human context to georeference the name. Due to naming variations between the travel journal and gazetteer and temporal disassociations between the time of the study and publish date of the travel journal, most POI names are identified as false negatives. The use of POI names for location analysis and temporal approximation narrows the geographic footprint of an event. This dissertation recognizes that the success of this study is dependent on identifying the means to name match and georeference POI names. This dissertation provides a foundation to support the inclusion of POI names in for GIR peer-reviewed studies, describes an implementation process and

`

research plan, and provides a mechanism to place POI names. The dissertation demonstrates that text representing placenames can be georeferenced when grouped and compared to other placenames in a travel journal.

Future Studies

Most peer-reviewed studies focus on toponyms, and as a result, their precision and recall rates are based mainly on toponyms. This dissertation focuses on both toponyms and POI names. Future studies should provide more focus on georeferencing POI names. Understanding the characteristics of a POI name can narrow the geographic footprint. Some questions related to POI studies are important to consider: What is the lifespan of a POI name? What prompts a person to record geographically sensitive data into a travel journal? What is the frequency of recording geographic information to one's journal? When do people record their experience in a travel journal? Is it during a trip, after a trip, or when something significant happens? Is there a relationship between the travel origination and the human agent's previous or current residency, employment, family or friend locations? How many unique toponyms exist in the United States to develop an authoritative list of placenames with minimal ambiguities? Currently, 25,044 unique toponyms were identified, but the count can go down when removing directional or descriptive words (e.g. River, Mountains, or Forest). How often are personal flashbacks or historical encounters mentioned within a travel journal? Socio-cultural studies and probabilities of trip destinations will become increasingly important as travel becomes more automated, such as allowing computers in

`

vehicles to determine the destinations based on personal travel behaviors and likelihood of destinations based on personal moods, time of day, and frequency of trips made at a certain time. Future studies of these questions will require the use of POI names. Access to POI names and toponyms will lead to an important venture in autonomous vehicles and modal destinations knowledge.

Multi-Modal Awareness

Autonomous vehicles (e.g. smart cars) are gaining popularity in the US. Forbes[34] Business Insider[35] listed safety, convenience, and energy efficiency as reasons to the popularity of driverless and smart cars. Elon Musk, Tesla CEO, predicted by the end of 2017 that the first self-driving car will be commercially available, although it will take decades before all cars will have self-driving capability[36]. The autonomous vehicle movement is spearheaded by major vehicle and data collection companies. Ford Automotive Company announced a five-year[37], $1 billion investment in development of artificial intelligence for hybrid autonomous vehicle[38]. Alphabet, Google's parent company, created a new department, Waymo[39], to lead the development of self-driving vehicles. Uber, a ridership program that relies on personal vehicles to act as a taxi and

[34] https://www.forbes.com/sites/modeledbehavior/2014/11/08/the-massive-economic-benefits-of-self-driving-cars/#66f0d0353273
[35] http://www.businessinsider.com/advantages-of-driverless-cars-2016-6/#roads-will-be-safer-1
[36] https://electrek.co/2017/02/13/tesla-elon-musk-all-new-cars-self-driving/
[37] https://www.washingtonpost.com/news/innovations/wp/2017/02/10/ford-to-invest-1-billion-in-artificial-intelligence-for-your-car/?utm_term=.e05129c09f94
[38] https://media.ford.com/content/fordmedia/fna/us/en/news/2016/12/28/ford-debuts-next-generation-fusion-hybrid-autonomous-development.html
[39] https://www.washingtonpost.com/news/innovations/wp/2016/12/13/google-is-one-step-closer-to-making-money-off-self-driving-cars/?utm_term=.208c1a4e3564

`

mobile devices to request and pay for the services, is currently testing the use of self-driving vehicles in Pittsburgh[40] and Arizona[41]. Finally, Honda have showcased a vehicle that has an "emotion engine[42]" designed to learn the driver's emotional pattern and act as a companion for the driver. Corporate entities are moving towards artificial intelligence to provide self-driving vehicles and an ability to learn to react to the driver's emotional needs. These needs may be useful in automating decisions making, narrowing options into a subset of likely choices, or offering location-based services to the driver's needs.

This dissertation is one approach to expand the role of autonomous vehicles to learn the driver's push and pull factors to a destination. The popularity of mobile and wearable devices, navigation and location-based services, and humans providing geographic-sensitive data provides the capability to learn the driver's destinations. The capability to learn the driver's behavior patterns and popular destinations is based on this dissertation's three properties (spatial, temporal, and human agent). Figure 45 specifies the attributes based on spatial, temporal, and human agent properties to provide a multi-modal awareness for vehicle autonomy. As more data is collected based on the driver's driving pattern the vehicle will learn the driver's destinations. The storage and maintenance of the data is a function that can be included in the overall geospatial cyberinfrastructure (Yang et al. 2010). Figure 46 shows the progression of data collection and storage from a human agent, and the capability for the multi-modal

[40] http://www.usatoday.com/story/tech/news/2016/09/22/uber-testing-self-driving-cars-san-francisco/90847962/#
[41] http://money.cnn.com/2017/03/27/technology/uber-san-francisco-self-driving-cars/
[42] https://www.washingtonpost.com/news/innovations/wp/2016/12/08/this-honda-concept-car-will-have-emotions-of-its-own/?utm_term=.bcbe644e238f

`

transportation to know the destination. Corporate entities have started the task of designing and implementing the first generation of autonomous vehicles, and the U.S. federal government has started drafting safety measures[43] to prepare for the inclusion of autonomous vehicles into society. Autonomous vehicles are progressing forward towards public use, and by incorporating the spatial, temporal and human agent properties defined by this dissertation the autonomous vehicles can determine the destination of a driver with little to no feedback from that driver.

[43] https://www.washingtonpost.com/local/trafficandcommuting/federal-officials-plan-aggressive-approach-to-driverless-cars/2016/09/19/3e78411e-7e92-11e6-8d0c-fb6c00c90481_story.html?utm_term=.a81b47b0223f

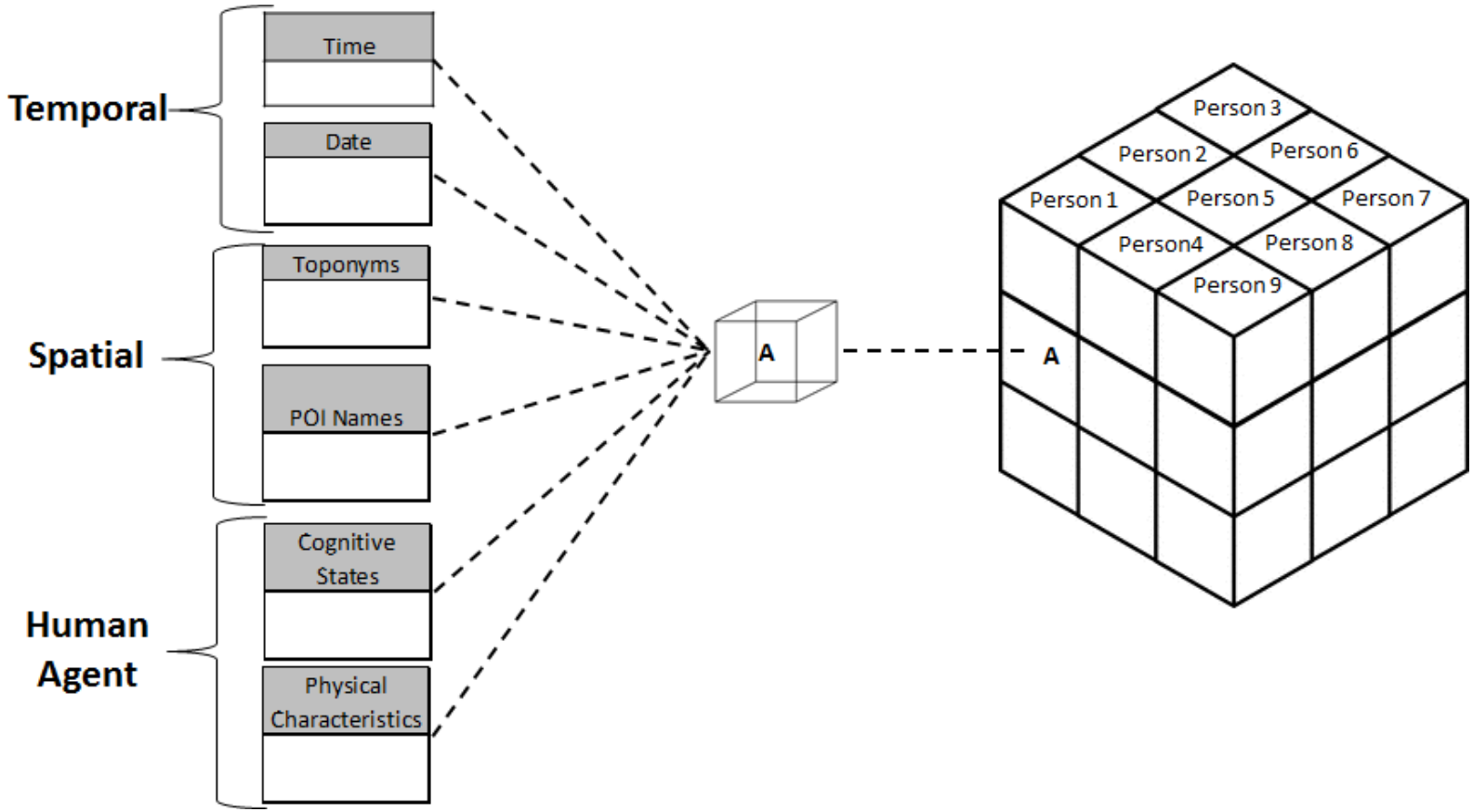

Figure 45 Cube diagram storing personal spatio-temporal locations based on dimensions.

`

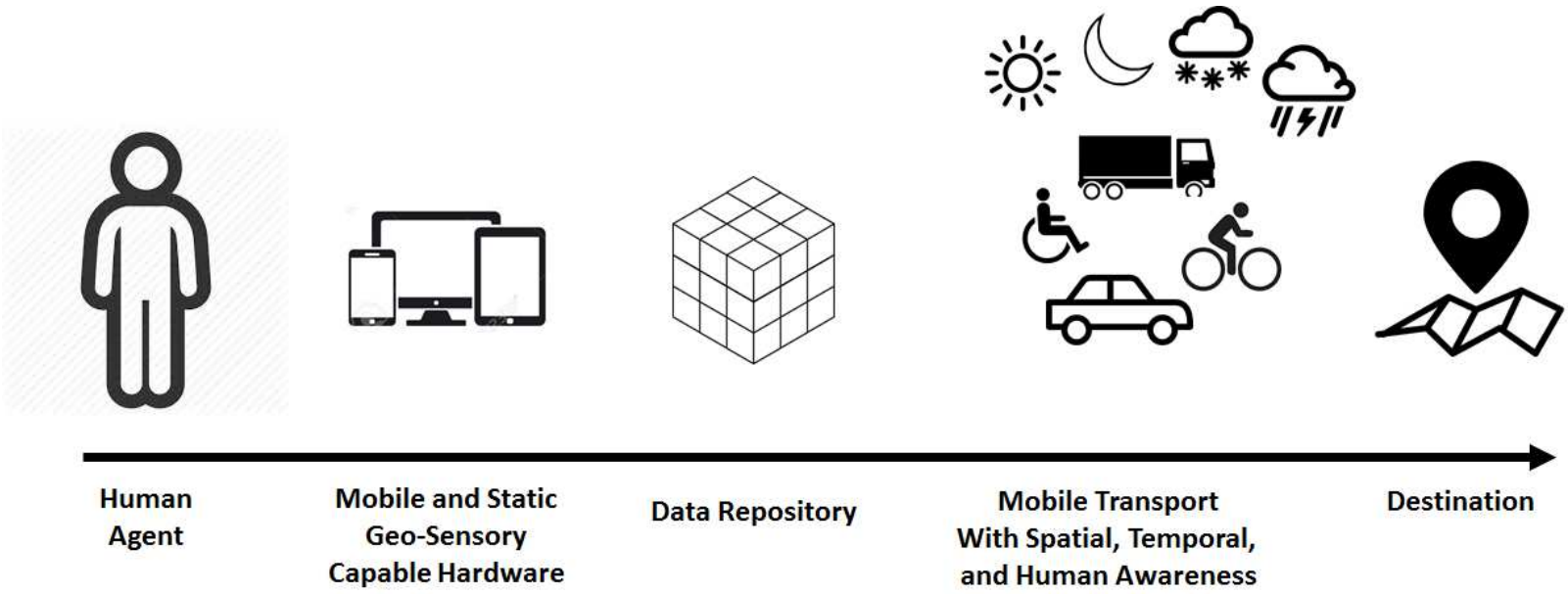

Figure 46 Destination knowledge provided to autonomous vehicles from geo-sensory devices.

`

APPENDIX A - Glossary

1. *Alternative names:* Names revised due to spelling variation or aliases.

2. *Annotations:* Semantic category of a text that is encapsulated by a token. The named entity recognizer and part-of-speech are examples of annotations

3. *Bag of words:* Unstructured, text-based documents such as travel journal in which focus is on the text and ignores grammar and word ordering.

4. *Corpus:* A known structured test documents used for statistical analysis.

5. *Gazetteers:* Metadata resource containing list of toponyms and their geographic information.

6. *Frequency Analysis:* Shows the number of occurrences.

7. *Geo-parsing:* Identifying ambiguous geographic text and assigning geographic information.

8. *Geo-referencing:* Assignment of geographic information to a placename. This is normally part of the geo-parsing stage which deals specifically with unstructured text files, but the geo-referencing stage is recognized as its own components for the proposal.

9. *Geographic Information Retrieval (GIR):* The overall process of geo-parsing unstructured text and indexing them for future retrieval.

10. *Geographic Information Systems (GIS):* A system designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data.

11. *Geographic Information Sciences (GISci):* The field of research that studies the theory and concepts that underpin GIS.

12. *Geographic scope:* The overall geographic location of a document.

13. *Heuristics:* A rule-based approach used to determine what constituted a POI name from a textual narrative. This approach determines token words that are often associated to POI names and stop words to encapsulate the names.

14. *K-Means Clustering:* Unsupervised learning algorithm that takes a known number of cluster, define the coordinate for each cluster group, and place the item whose own geographic reference is closest to the cluster group centroid.

`

15. *Lexical Ambiguities:* Words that can be used to express two or more meanings. Geographic names are often confused with other geographic words or non-geographic words.

16. *Named Entity Recognition (NER):* Labelling of text entities representing people and things

17. *Natural Language Processing (NLP):* Information extraction that locates and classifies named entities into a pre-defined group such as name of person, location, or organization.

18. *Taxonomy Nodes:* An element within a taxonomy tree normally representing a toponym or a POI name.

19. *Tokens: A semantic encapsulation of a text that also contains annotations such as the named entity recognizer which defines a text as a location and part-of-speech which describe if a text is a proper noun.*

20. *Points-of-Interest (POI):* Attractions or designations often characterized as human-developments.

21. *Precision:* Calculates the quality of the existing data. Asks "of all the toponyms and POI names collected, what is the ratio or percentage of those items being true positives". Precision are positive predictive value as it measures the ratio of all positive values divided by the sum of the positive and false positive values.

22. *Recall:* Calculates the quality of the data collected. Ask "of all the possible candidates stored in the travel journal, how many were actually identified as true positives".

23. *Semantic Ambiguities:* A word is interpreted more than one way due to the ambiguous sentence structure.

24. *Taxonomy tree*: A hierarchy tree consisting of toponyms and terminating with a POI name. The taxonomy tree will contain nodes arranging the toponyms in hierarchy order [State→City→POI name]. The highest-level node will be the geographic scope, and the lowest will contain the POI name or the subsequent toponym.

25. *Toponyms:* An administrative or political placename defined by an authoritative source and often found in traditional gazetteers.

26. *Toponym resolution:* The area based on the lowest toponym described in the taxonomy tree. As the scale of the toponym increases so does the resolution.

`

APPENDIX B – Technical Requirements

The software and application requirements include:

1. Text File to store each chapter.
2. Stanford NLP (geoparsing application).
3. Bill Bryson's *The lost continent: travels in small-town America*. Travel Journal (Data source).
4. Python 2.7 (program to provide georeferencing, frequency, and clustering engine).
5. Geonames Gazetteer (metadata).
6. Nominatim locator (metadata).
7. JavaScript and Leaflet API (web map).
8. ArcGIS (reviewer application).

The hardware:

1. Microsoft Surface Pro 3.
2. Intel Core i5-4300U CPU 1.90GHz.
3. 4 GB RAM.
4. 64-bit Microsoft Window 10 Pro Operating System.

`

APPENDIX C – Stanford NLP Script to Parse the Text in the Travel Journal

```
java -cp "*" -Xmx2g edu.stanford.nlp.pipeline.StanfordCoreNLP -annotators
tokenize,ssplit,pos,lemma,ner,parse,dcoref -file
C:\geo_parsing\corpus\SmallTown_1.txt
```

APPENDIX D1 – Wireframe Process to Identify the Travel Journal Origination

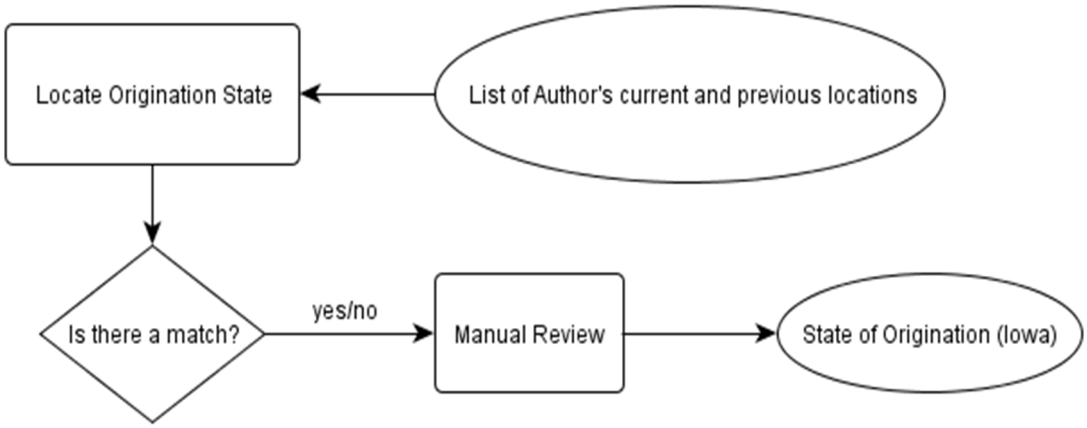

Figure 47 Location of the starting/origination state for Chapter 1

APPENDIX D2 – Wireframe Process of the Frequency Analysis

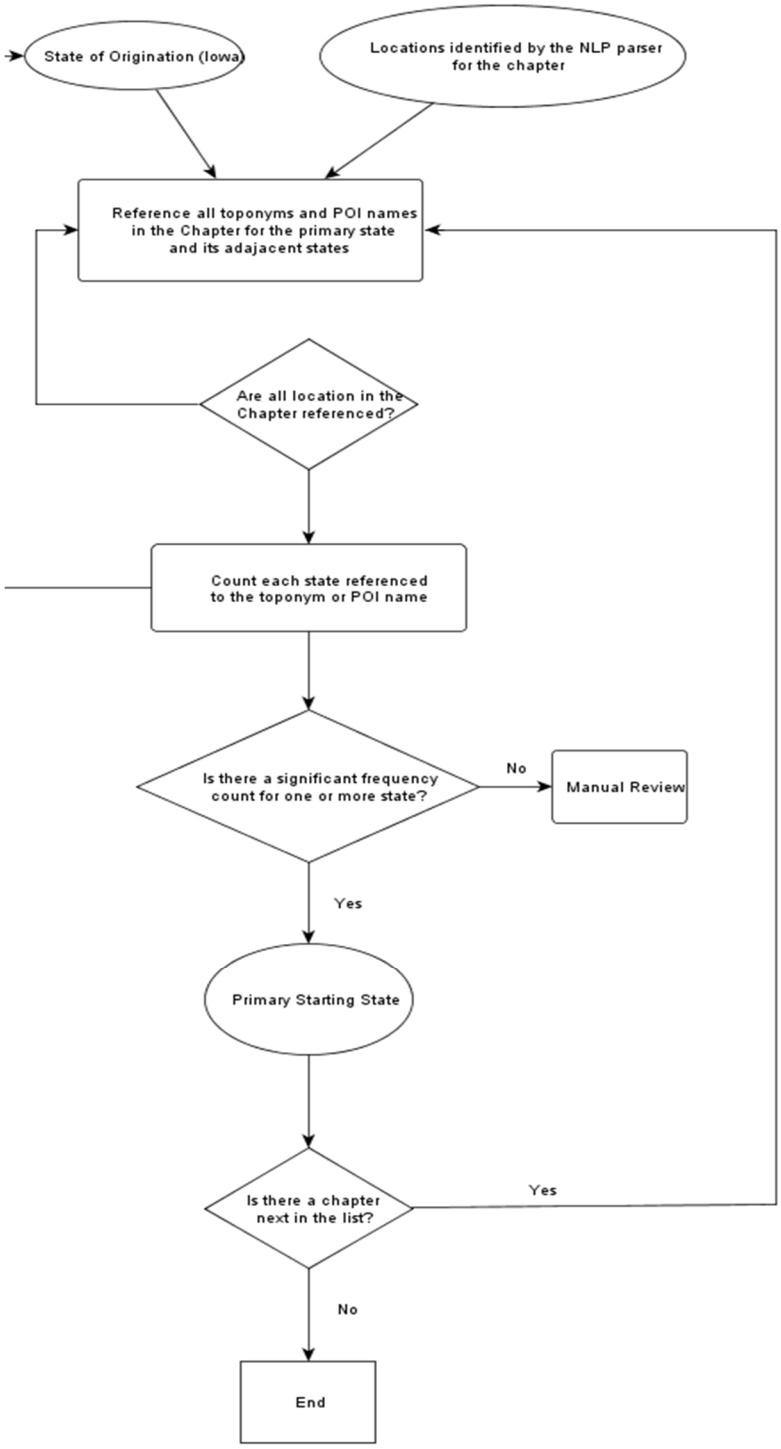

Figure 48 Frequency analysis process.

APPENDIX D3 – Wireframe Process of the Proximity Clustering Analysis

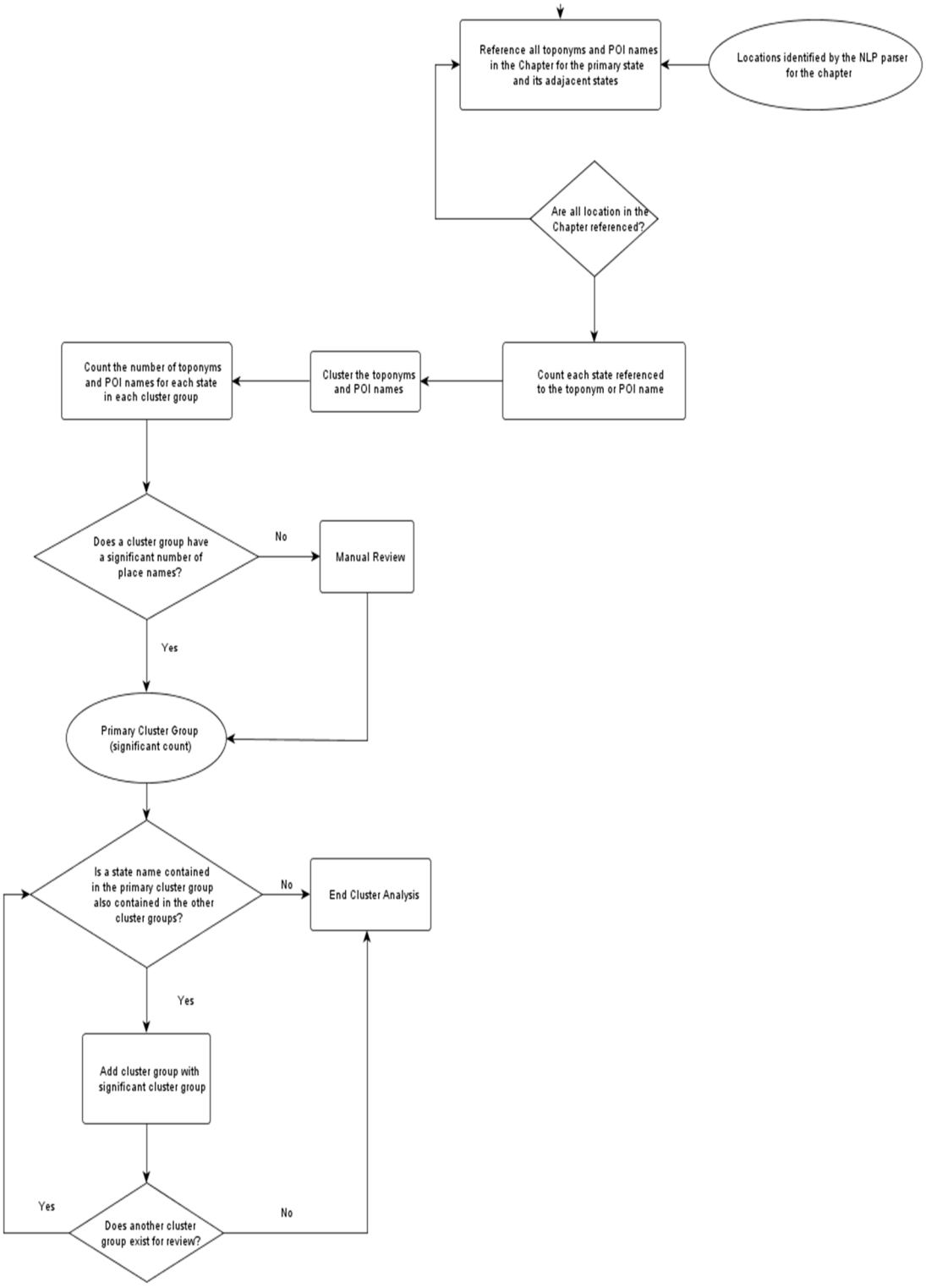

Figure 49 Cluster analysis process.

REFERENCES

Achen, C. H. (2002). Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science*, *5*(1), 423-450.

Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: geotagging web content. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04 (pp. 273–280). New York, NY, USA: ACM.

Andogah, G., Bouma, G., and Nerbonne, J. (2012). Every document has a geographical scope. Data and Knowledge Engineering, (0).

Andreu, L., Kozak, M., Avci, N., and Cifter, N. (2006). Market segmentation by motivations to travel: British tourists visiting Turkey. *Journal of Travel and Tourism Marketing*, *19*(1), 1-14.

Bae, B. C., & Young, R. M. (2008, November). A use of flashback and foreshadowing for surprise arousal in narrative using a plan-based approach. In *Joint International Conference on Interactive Digital Storytelling* (pp. 156-167). Springer Berlin Heidelberg.

Banu, Y. Sophiya, Y. Soniya Banu, and V. V. Karthikeyan. (2013). A Novel Approach For Georeferenced Data Analysis Using Hard Clustering Algorithm. *International Journal of Research in Engineering and Technology. 2*(5), 783-786.

Bargh, J. A., and McKenna, K. Y. (2004). The Internet and social life. *Annu. Rev. Psychol.*, *55*, 573-590.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. American Psychologist, 32(7), 513–531.

Bryson, B. (1989). *The lost continent: travels in small-town America*. New York: Harper and Row.

Bulson, E. (2007). *Novels, maps, modernity: the spatial imagination, 1850-2000*. Routledge.

Buscaldi, D. (2011). Approaches to disambiguating toponyms. *SIGSPATIAL Special*, *3*(2), 16-19.

Calì, D., Condorelli, A., Papa, S., Rata, M., and Zagarella, L. (2011). Improving intelligence through use of Natural Language Processing. A comparison between NLP interfaces and traditional visual GIS interfaces. *Procedia Computer Science*, *5*, 920-925.

Cave, April (2016). Reconstructing Paths of Travel Based on Location References in Unstructured Text. George Mason University.

Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J. and Waters, N. M. 2016. Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. International Journal of Remote Sensing, 37:1, 100-124, DOI: 10.1080/01431161.2015.1117684

Cestra, G., Liguori, G., and Clementini, E. (2011). MyTravel: a geo-referenced social-oriented web 2.0 application. In Computational Science and Its Applications-ICCSA 2011 (pp. 225-236). Springer Berlin Heidelberg.

Coleman, D. J., Georgiadou, Y., and Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. International Journal of Spatial Data Infrastructures Research, 4(1), 332-358.

Conrad, C. C., and Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, *176*(1-4), 273-291.

Crane, R., and Crepeau, R. (1998). Does neighborhood design influence travel?: A behavioral analysis of travel diary and GIS data1. Transportation Research Part D: Transport and Environment, 3(4), 225–238.

Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources.

Drymonas, E., Efentakis, A., and Pfoser, D. (2011, September). Opinion mapping travelblogs. In *Proceedings of Terra Cognita workshop (in conjunction with the 10th international semantic web conference)* (pp. 23-36).

Elwood, S. (2008). Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal*, *72*(3-4), 173-183.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. doi:10.1080/00045608.2011.595657

Gatewood, J. B. (1983). Loose Talk: Linguistic Competence and Recognition Ability. *American Anthropologist*, *85*(2), 378–387.

Gelernter, J., and Balaji, S. (2013). An algorithm for local geoparsing of microtext. GeoInformatica, 17(4), 635–667.

Gelernter, J., and Mushegian, N. (2011). Geo-parsing Messages from Microtext. *Transactions in GIS*, *15*(6), 753-773.

Gervais, Marc, et al. "Data quality issues and geographic knowledge discovery." *Geographic data mining and knowledge discovery* (2009): 99-115.

Gey, F., Larson, R., Kando, N., Machado, J., and Sakai, T. (2010, June). NTCIR-GeoTime overview: Evaluating geographic and temporal search. In *NTCIR* (Vol. 10, pp. 147-153).

Goodchild, M., and Hill, L. L. (2006). Digital Gazetteer and Practice Workshop, Summary Report. eScholarship.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69*(4), 211-221.

Goodchild, M. F., Kyriakidis, P., Rice, M., & Schneider, P. (2005). Spatial Webs, Final Report and Position Papers. https://escholarship.org/uc/item/46z721n2

Goodchild, Michael F., and J. Alan Glennon. "Crowdsourcing geographic information for disaster response: a research frontier." *International Journal of Digital Earth* 3.3 (2010): 231-241.

Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., and Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875–3889.

Guo, Q., Li, W., Liu, Y., and Tong, D. (2011). Predicting potential distributions of geographic events using one-class data: concepts and methods. International Journal of Geographical Information Science, 25(10), 1697–1715.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On Clustering Validation Techniques. Journal of Intelligent Information Systems, 17(2–3), 107–145. http://doi.org/10.1023/A:1012801612483

Harvey, D. (2001). *Spaces of capital: Towards a critical geography*. Routledge.

Hertel, G., Niedner, S., and Herrmann, S. (2003). Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research policy*, *32*(7), 1159-1177.

Hill, L. L. (2009). *Georeferencing: The Geographic Associations of Information*. The MIT Press.

Hirsh, J. B., and Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of research in personality*, *43*(3), 524-527.

Holt-Jensen, Arild (2009) *Geography: history and concepts: a student's guide*, 4th editin. London: SAGE.

Hsu, C. L., and Lin, J. C. C. (2008). Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information and management*, *45*(1), 65-74.

Jakobson, R. (1960). Linguistics and poetics. In *Style in language* (pp. 350-377). MA: MIT Press.

Johnson, D. W. (1917). The Role of Political Boundaries. *Geographical Review*, *4*(3), 208–213.

Johnson, T. J., and Kaye, B. K. (2002). Webelievability: A path model examining how convenience and reliance predict online credibility. *Journalism and Mass Communication Quarterly*, *79*(3), 619-642.

Jones, C. B., Purves, R. S., Clough, P. D., and Joho, H. (2008). Modelling vague places with knowledge from the Web. *International Journal of Geographical Information Science*, *22*(10), 1045–1065.

Jönsson, C., and Devonish, D. (2008). Does Nationality, Gender, and Age Affect Travel Motivation? a Case of Visitors to The Caribbean Island of Barbados. Journal of Travel and Tourism Marketing, 25(3–4), 398–408.

Keßler, C., Janowicz, K., and Bishr, M. (2009). An Agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (pp. 91–100). New York, NY, USA.

Kemp, K. (2008). *Encyclopedia of geographic information science*. Sage

Ketchen Jr, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. Strategic management journal, 441-458.

Larson, R. R. (1996). Geographic information retrieval and spatial browsing. Retrieved from http://hdl.handle.net/2142/416

Larson, R. R. (2011). Ranking approaches for GIR. SIGSPATIAL Special, 3(2), 37–41.
Lasersohn, P. (1999). Pragmatic Halos. Language, 75(3), 522–551.

Lay, J. G., Chen, Y. W., and Yap, K. H. (2010). Geographic reality versus imagination in Taiwan's historical maps. *The Cartographic Journal*, *47*(2), 180-189.

Leidner, J.L., 2017. Georeferencing: From Texts to Maps. In: Richardson, D., Castree, N., Goodchild, M. F., Kobayashi, A., Liu, W. and Marston, R. (Eds.), International Encyclopedia of Geography: People, the Earth, Environment, and Technology. Wiley: New York.

Leidner, J. L., and Lieberman, M. D. (2011). Detecting geographical references in the form of placenames and associated spatial natural language. *SIGSPATIAL Special*, *3*(2), 5–11.

Lewis, D. D., and Jones, K. S. (1996). Natural language processing for information retrieval. *Communications of the ACM*, *39*(1), 92-101.

Li, C., & Sun, A. (2014, July). Fine-grained location extraction from tweets with temporal awareness. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 43-52). ACM.

Loukas, A., and Vasiliades, L. (2004). Probabilistic analysis of drought spatiotemporal characteristics inThessaly region, Greece. Natural Hazards and Earth System Science, 4(5/6), 719–731.

Longley, P. A., Goodchild, M. F., Maguire, D. J., & Rhind, D. W. (2015). Geographic information science and systems. John Wiley & Sons.

Mansfeld, Y. (1992). From motivation to actual travel. *Annals of tourism research*, *19*(3), 399-419.
McIntosh, J., and Yuan, M. (2005). Assessing Similarity of Geographic Processes and Events. Transactions in GIS, 9(2), 223–245.

Medina, R.M., Cervone, G. and Waters, N.M., 2017. Characterizing and predicting traffic accidents in extreme weather environments. The Professional Geographer, 69(1), pp.126-137

Michalowski, M., and Knoblock, C. A. (2005, July). A constraint satisfaction approach to geospatial reasoning. In *AAAI* (Vol. 2005, pp. 423-429).

Miller, Harvey J. 1999. “Measuring Space-Time Accessibility Benefits within Transportation Networks: Basic Theory and Computational Procedures.” *Geographical Analysis* 31 (1): 1–26.
Moncla, L., Gaio, M., & Mustiere, S. (2014, September). Automatic itinerary reconstruction from texts. In International Conference on Geographic Information Science (pp. 253-267). Springer International Publishing.

Mukařovský, J. (2014). Standard language and poetic language. *Chapters from the history of Czech functional linguistics*.
Nadkarni, P. M., Ohno-Machado, L., and Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Peregrino, F., Tomás, D., and Pascual, F. (2012). Question Answering and Multi-search Engines in Geo-Temporal Information Retrieval. In A. Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science (Vol. 7182, pp. 342–352). Springer Berlin / Heidelberg.

Pradhan, B., and Youssef, A. M. (2009). Manifestation of remote sensing data and GIS on landslide hazard analysis using spatial-based statistical models. Arabian Journal of Geosciences, 3(3), 319–326.

Pritchard, A., and Morgan, N. J. (2000). Constructing tourism landscapes - gender, sexuality and space. Tourism Geographies, 2(2), 115–139.

Radding, L., and Western, J. (2010). What’s in a Name? Linguistics, Geography, and Toponyms. *Geographical Review*, *100*(3), 394-412.

Rauch, E., Bukatin, M., and Baker, K. (2003). A confidence-based framework for disambiguating geographic terms. *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references - Volume 1*, HLT-NAACL-GEOREF ’03 (pp. 50–54). Stroudsburg, PA, USA: Association for Computational Linguistics.

Rice, M. T., Aburizaiza, A. O., Jacobson, R. D., Shore, B. M., and Paez, F. I. (2012a). Supporting Accessibility for Blind and Vision-impaired People With a Localized Gazetteer and Open Source Geotechnology. *Transactions in GIS*, *16*(2), 177-190.

Rice, M. T., Hammill, W. C., Aburizaiza, A. O., Schwarz, S., and Jacobson, R. D. (2011). Integrating user-contributed geospatial data with assistive geotechnology

using a localized gazetteer. In *Advances in Cartography and GIScience. Volume 1* (pp. 279-291). Springer Berlin Heidelberg.

Rice, M. T., Paez, F. I., Mulhollen, A. P., Shore, B. M., & Caldwell, D. R. (2012b). Crowdsourced Geospatial Data: A report on the emerging phenomena of crowdsourced and user-generated geospatial data (Annual No. AA10-4733). Fairfax, VA: George Mason University. http://www.dtic.mil/dtic/tr/fulltext/u2/a576607.pdf. Accessed 5 July 2013

Rice, M. T., Jacobson, R. D., Caldwell, D. R., McDermott, S. D., Paez, F. I., Aburizaiza, A. O., … Qin, H. (2013a). Crowdsourcing techniques for augmenting traditional accessibility maps with transitory obstacle information. *Cartography and Geographic Information Science*, *40*(3), 210–219.

Rice, M. T., Curtin, K. M., Paez, F. I., Seitz, C. R., & Qin, H. (2013b). Crowdsourcing to Support Navigation for the Disabled: A Report on the Motivations, Design, Creation and Assessment of a Testbed Environment for Accessibility (US Army Corps of Engineers, Engineer Research and Development Center, U.S. Army Topographic Engineering Center Technical Report, Data Level Enterprise Tools Workgroup No. BAA: #AA10-4733, Contract: # W9132V-11-P-0011) (pp. 1–62). Fairfax, VA: George Mason University. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA588474

Rice, M. T., Paez, F. I., Rice, R. M., Ong, E. W., Qin, H., Seitz, C. R., et al. (2014). Quality Assessment and Accessibility Applications of Crowdsourced Geospatial Data: A report on the development and extension of the George Mason University Geocrowdsourcing Testbed (Annual No. BAA: #AA10-4733, Contract: # W9132V-11-P-0011) (p. 91). Fairfax, VA: George Mason University. http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=ADA615952

Rice, M. T., Paez, F. I., Rice, R. M., Ong, E. W., Qing, H., Seitz, C. R., … Medina, R. M. (2014). Quality assessment and accessibility applications of crowdsourced geospatial data. Geospatial Research Laboratory - U.S. Army Engineer Research and Development Center - U.S. Army Corps of Engineers.

Rice, M. T., Curtin, K. M., Pfoser, D., Rice, R. M., Fuhrmann, S., Qin, H., … Paez, F. I. (2015). *Social Moderation and Dynamic Elements in Crowdsourced Geospatial Data: A Report on Quality Assessment, Dynamic Extensions and Mobile Device Engagement in the George Mason University Geocrowdsourcing Testbed*. George Mason University Fairfax United States. Retrieved from http://www.dtic.mil/cgi-bin/GetTRDoc?Location=U2&doc=GetTRDoc.pdf&AD=AD1001943Rodgers, R.

(2015). *A Statistical Comparison of Sidewalk Slopes Derived From Multi-resolution Digital Elevation Models in Support of Accessibility*. George Mason University.

Rice, R. M., Aburizaiza, A. O., Rice, M. T., & Qin, H. (2016). Position Validation in Crowdsourced Accessibility Mapping. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *51*(2), 55–66.

Sack, R.D. (1997) Homo Geographicus a Framework for Action, Awareness, and Moral Concern. Baltimore, Maryland: Johns Hopkins University Press.

Schmallegger, D., and Carson, D. (2008). Blogs in tourism: Changing approaches to information exchange. *Journal of vacation marketing*, *14*(2), 99-110.

Seeger, C. J. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. GeoJournal, 72(3–4), 199–213.

Silvis, J., Niemeier, D., and D'Souza, R. (2006, August). Social networks and travel behavior: report from an integrated travel diary. In *11th International Conference on Travel Behaviour Reserach, Kyoto*.

Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Cardoso, N. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, *30*(4), 378–399.

Smith, D., and Crane, G. (2001). Disambiguating Geographic Names in a Historical Digital Library. In P. Constantopoulos and I. Sølvberg (Eds.), *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science (Vol. 2163, pp. 127–136). Springer Berlin / Heidelberg.

Smith, T. R., and Frew, J. (1995). Alexandria Digital Library. *Commun. ACM*, *38*(4), 61–62.

Sönmez, S. F., and Graefe, A. R. (1998). Influence of terrorism risk on foreign tourism decisions. *Annals of Tourism Research, 25*(1), 112-144.

Sperber, D., and Wilson, D. (1985). Loose Talk. *Proceedings of the Aristotelian Society*, *86*, 153–171.

Stefanidis, A., Crooks, A., and Radzikowski, J. (2013). Harvesting ambient geospatial information from social media feeds. *GeoJournal*, *78*(2), 319-338.

Swain, M. B. (1995). Gender in tourism. Annals of Tourism Research, 22(2), 247–266.

Teo, T. S., Lim, V. K., and Lai, R. Y. (1999). Intrinsic and extrinsic motivation in Internet usage. *Omega*, *27*(1), 25-37.

Thompson, J. (2011). National Geographic Traveler: Washington, DC, 4th edition (4th ed.). National Geographic.

Timothy, D. J. (1995). Political boundaries and tourism: borders as tourist attractions. *Tourism Management*, *16*(7), 525–532.

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234–240.

Um, S., and Crompton, J. L. (1990). Attitude determinants in tourism destination choice. Annals of Tourism Research, 17(3), 432–448.

Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In *Advances in Spatial and Temporal Databases* (pp. 218-235). Springer Berlin Heidelberg.

Varzi, A. C. (2001). Vagueness in geography. Philosophy and Geography, 4(1), 49–65.

Verma, S., Vieweg, S., Corvey, W. J., Palen, L., Martin, J. H., Palmer, M., ... and Anderson, K. M. (2011, July). Natural Language Processing to the Rescue? Extracting" Situational Awareness" Tweets During Mass Emergency. In *ICWSM*.

Wacholder, N., Ravin, Y., and Choi, M. (1997). Disambiguation of proper names in text. *Proceedings of the fifth conference on Applied natural language processing*, ANLC '97 (pp. 202–208). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005). Detecting geographic locations from web resources. In *Proceedings of the 2005 Workshop on Geographic Information Retrieval* (pp. 17–24). New York, NY, USA: ACM.

Waters, N. 2013. Social Network Analysis. In Fischer M.M., Nijkamp P. (Eds) Handbook of Regional Science, Ch. 38, pp. 725-740. Springer: Heidelberg, New York, Dordrecht, London

Waters, N. M. 2017. Tobler's First Law of Geography. In: Richardson, D., Castree, N., Goodchild, M. F., Kobayashi, A., Liu, W. and Marston, R. (Eds.), International Encyclopedia of Geography: People, the Earth, Environment, and Technology. Wiley: New York.

Wolf, J., Guensler, R., and Bachman, W. (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board*, (1768), 125-134.

Wood, Denis (2010) Rethinking the Power of Maps. New York: The Guilford Press.

Woodruff, A. G., and Plaunt, C. (1994b). GIPSY: Geo-referenced Information Processing System. Journal of the American Society for Information Science, 45(9), 645-655.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, *44*(3), 363-373.

Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., ... & Fay, D. (2011). Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? International Journal of Digital Earth, 4(4), 305-329.

Zhang, Q., Jin, P., Lin, S., Yue, L.: Extracting Focused Locations for Web Pages. In: Wang, L., Jiang, J., Lu, J., Hong, L., Liu, B. (eds.) WAIM 2011 Workshops. LNCS, vol. 7142, pp. 76–89. Springer, Heidelberg (2012).

Zheng, Y., Zhang, L., Xie, X., and Ma, W.-Y. (2009). Mining Interesting Locations and Travel Sequences from GPS Trajectories. In Proceedings of the 18th International Conference on World Wide Web (pp. 791–800). New York, NY, USA: ACM.

Zook, M., Graham, M., Shelton, T., and Gorman, S. (2010). Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake. *World Medical and Health Policy*, *2*(2), 7-33.

BIOGRAPHY

Scott D. McDermott is a GIS administrator working as a federal contractor in Washington, D.C. His study of interest is in evaluating various socio-economic implications and its geographic impacts within urban and regional geographic areas; as well as advancing the use of GIS. Scott currently lives in Washington, D.C.