RECONSTRUCTING PATHS OF TRAVEL BASED ON LOCATION REFERENCES IN UNSTRUCTURED TEXT

by

April Cave
A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

Dr. Matthew Rice, Thesis Director
Dr. Aire Croitoru, Committee Member
Dr. Sven Fuhrmann, Committee Member
Dr. Anthony Stefanidis, Department Chairperson
Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
Dr. Peggy Agouris, Dean, College of Science

Date: April 20, 2018

Spring Semester 2016
George Mason University
Fairfax, VA
Reconstructing Paths of Travel Based on Location References in Unstructured Text

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University.

by

April Cave
Bachelor of Science
George Mason University, 2012

Director: Matthew Rice, Associate Professor
Department of Geography and Geoinformation Science

Spring Semester 2016
George Mason University
Fairfax, VA
This work is licensed under a creative commons attribution-noderivs 3.0 unported license.
DEDICATION

I would like to dedicate this project to my family. In particular, my mom, my aunt Patti and my little brother, George.

My mom, Pam, has provided support, encouragement and resources throughout my entire life. She has been there through the most difficult parts of my life and never said a negative word about my educational or career pursuits. I would not have made it through these last ten years or been able to finish undergrad, let alone a masters program without her help.

My aunt Patti has also provided support, encouragement and resources throughout everything. She also acted as a soundboard for many of my ideas and during the progress of this project. She was, and will continue to be, one of my biggest cheerleaders.

My little brother, George, is the most important person in my life. He has been my inspiration to achieve beyond expectations since he was born. It is very important to me to show him that it is possible to achieve big things even though life may throw great obstacles his way. He was always curious in what I was working on and asked a lot of questions. He always seemed genuinely interested. He was encouraging when I got frustrated and it has never meant more than when he said he was proud of me.
ACKNOWLEDGEMENTS

There are several people I would like to thank for aiding the completion of this project.

First I would like to thank my thesis director, Dr. Matthew Rice. We first met while I was an undergrad at GMU in an introductory GIS course and was encouraging when I contacted him about my interest in pursuing an MS in Geoinformatics and Geospatial Intelligence. He was willing to have long discussions about ideas and provided valuable feedback throughout the creation of this thesis and my graduate education. He was also largely responsible for helping me to obtain space and adaptive equipment that enabled me to complete my schoolwork.

Next, I would like to thank the rest of my thesis committee, Dr. Arie Croitoru and Dr. Sven Fhurman for their attention and feedback.

I would like to thank my entire committee for hanging in when I had several unforeseen obstacles occur in my life.

Finally, I would like to thank Debbie Hutton and Ron Mahabir, who helped to carryout the space and adaptive equipment requests that were made.
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................................................................VII

LIST OF ABBREVIATIONS ....................................................................................................VIII

ABSTRACT ............................................................................................................................ IX

INTRODUCTION .................................................................................................................... 1
  JULES VERNE .................................................................................................................... 3
  Who was Jules Verne? ....................................................................................................... 3
  Relevant characteristics of Verne’s stories ....................................................................... 3
  THESIS ORGANIZATION ............................................................................................... 7

CONCEPTUAL FRAMEWORK ............................................................................................... 8
  LITERATURE REVIEW ...................................................................................................... 8
  STATEMENT OF SCIENTIFIC RESEARCH QUESTIONS .................................................. 13

DATA & METHODOLOGY ...................................................................................................... 14
  DATA SOURCES ................................................................................................................ 14
  DATA COLLECTION AND PROCESSING .......................................................................... 16
  Location References Data Table ....................................................................................... 17
  Retrieving Location References from the Text ................................................................. 21
  Determining Relevance of Location References ............................................................. 26
  Geocodinging Location References ............................................................................... 30

RESULTS & DISCUSSION ..................................................................................................... 47
  HOW THIS PROCESS CAN BE USED BY THE USIC AND LAW ENFORCEMENT .......... 64

CONCLUSIONS & FUTURE WORK ...................................................................................... 67
  CONCLUSIONS ................................................................................................................ 67
  FUTURE RESEARCH ....................................................................................................... 68

REFERENCES ....................................................................................................................... 71

BIOGRAPHY ......................................................................................................................... 75
LIST OF FIGURES

Figure 1: Fields contained within the location references data table ........................................ 17
Figure 2: Example of page from source text. a) chapter number; b) page number; c) paragraph count (Hilbert, 2012, p. 689) ................................................................. 23
Figure 3: Example of a location reference within a sentence containing a semi-colon ... 24
Figure 4: Example of an indirect reference (Hilbert, 2012, p. 593) ........................................ 25
Figure 5: Exception to the overall relevance rule (Hilbert, 2012, p. 577) .............................. 29
Figure 6: GNS search: a) Include Historical Records; b) contains; c) Enter location reference of interest; d) Search Database; e) Additional search requirements
(Geonames.nga.mil [GNS], 2015) ............................................................................................. 31
Figure 7: GNS search results: a) number of results; b) List of results; c) New results with more search requirements; d) latitude and longitude; e) alternate name/spelling
(Geonames.nga.mil [GNS], 2015) ............................................................................................. 32
Figure 8: TGN search: a) Location reference; b) search; c) More search requirements
(“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016) ....................... 33
Figure 9: Initial TGN search results: a) Alphabetical order; b) number of results ("Getty Thesaurus of Geographic Names (Getty Research Institute)", 2016) ......................... 34
Figure 10: Second set of TGN results: a) Alphabetical order; b) fewer results (“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016) ......................... 34
Figure 11: TGN London result: a) Select appropriate result; b) latitude and longitude in decimal degrees; c) alternate names/spellings (“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016) .................................................................................. 35
Figure 12: Google Earth search: a) Enter location; b) Select Search (Google Earth, 2016) ............................................................................................................................................ 36
Figure 13: Add point in GE: a) Select add new point; b) fill out location information; c) latitude and longitude (Google Earth, 2016) ......................................................... 37
Figure 14: Multi-ring buffer used to determine indirect references to locations outside of Allahabad (Google Earth, 2016) (ArcGIS 10.3.1 for Desktop, 2016) .................. 39
Figure 15: Finding Milligaum (“Great Indian Peninsula Railway”, 2016) ............................. 41
Figure 16: Malegaon (Google Earth, 2016) ............................................................................. 42
Figure 17: the University of Western Ontario, AKA Western University ("Western University", 2016) ................................................................................................................. 45
Figure 18: Percentage of relevance for the 1,337 references collected .................................. 48
Figure 19: Reference Type ........................................................................................................ 50
Figure 20: Ambiguity Type ...................................................................................................... 52
Figure 21: Burdivan (Google Earth, 2016) ............................................................................ 53
Figure 22: All 1,337 locations ................................................................................................. 55
Figure 23: 202 relevant points ............................................................................................... 55
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department of Defense</td>
<td>DoD</td>
</tr>
<tr>
<td>Digital South Asia Library</td>
<td>DSAL</td>
</tr>
<tr>
<td>Geographic Information Retrieval</td>
<td>GIR</td>
</tr>
<tr>
<td>Geographic Information System</td>
<td>GIS</td>
</tr>
<tr>
<td>GEONet Names Server</td>
<td>GNS</td>
</tr>
<tr>
<td>Getty Thesaurus of Geographic Names Online</td>
<td>TGN</td>
</tr>
<tr>
<td>Google Earth</td>
<td>GE</td>
</tr>
<tr>
<td>Keyhole Markup Language</td>
<td>KML</td>
</tr>
<tr>
<td>Law Enforcement</td>
<td>LE</td>
</tr>
<tr>
<td>National Football League</td>
<td>NFL</td>
</tr>
<tr>
<td>National Geospatial-Intelligence Agency</td>
<td>NGA</td>
</tr>
<tr>
<td>Not Applicable</td>
<td>NA</td>
</tr>
<tr>
<td>United States Geological Survey</td>
<td>USGS</td>
</tr>
<tr>
<td>United States Intelligence Community</td>
<td>USIC</td>
</tr>
</tbody>
</table>
ABSTRACT

RECONSTRUCTING PATHS OF TRAVEL BASED ON LOCATION REFERENCES IN UNSTRUCTURED TEXT

April Cave, M.S.
George Mason University, 2016
Thesis Director: Dr. Matthew Rice

The process of providing spatial context for location information within text has been explored by researchers in the domains of geographic information retrieval (GIR), geosocial media analysis, computational linguistics, and the digital humanities. This thesis will explore the topic of geographic information retrieval from text, using the stories of Jules Verne, which are fictional, but set in the real world. Just as location references, including place names and points of interest, within the text of Verne’s stories can be geoparsed and geocoded to be placed on a map and visualize the journey, so can location references within the text of intelligence gathered by the intelligence community and law enforcement. Once location references are extracted, they can be verified using outside sources and, by considering their spatial, temporal and thematic relevance, their significance determined. All verified, relevant locations can be placed on a map. This information can then be used to reconstruct the movement of a person of interest. The
methods in this thesis are used to document the positions of a subject through travels eastward across three continents.
INTRODUCTION

One of the goals of the United States Intelligence Community (USIC) and Law Enforcement (LE) is to prevent crime or acts of terrorism from occurring. Knowing where these and related events leading up to them (such as planning) are happening and being able to reconstruct the movements of persons of interest in these cases is vital to being able to prevent them.

The USIC and LE use multiple sources of intelligence in order to obtain this information. This includes traditional geospatial data, as well as text documents, such as emails, letters or text messages, or transcribed dialog, such as phone calls and recorded conversations. Finding the physical source of some of this can be done fairly easily. For instance, tracing the IP address from which an email was sent or the geotags embedded in pictures. However, it is more difficult to provide spatial context for the non-specific location information that may be contained within the text itself.

This process is experimental, and has been explored by researchers in the domains of geographic information retrieval (GIR), geosocial media analysis, computational linguistics, and the digital humanities. An interesting avenue for this research is the application of geographic information extraction techniques to travel narratives, historical fiction, and stories from authors such as Jules Verne, who made extensive use of geographic terms, geographic locations, and geographic settings for his science fiction.
This thesis will explore the topic of geographic information retrieval from text, using the stories of Jules Verne, which are fictional, but set in the real world. Just as location references, including place names and points of interest, within the text of Verne’s stories can be geoparsed and georeferenced to be placed on a map and visualize the journey, so can location references within the text of gathered intelligence. Verne’s stories possess several characteristics that are applicable to intelligence work: 1) Verne was from France; 2) his novels were originally published in French and translated to other languages, including English; 3) his novels were all written before 1905; and, 4) his stories include references to events that are not relevant to the current event(s). In the real world, regions use different references for the same feature, meanings can be ‘lost in translation’ or change over time, and include statements irrelevant to the current incident. All of these issues impact geospatial intelligence in specific ways. An incorrect translation may lead to the identification of the incorrect location. In addition, inclusion of events that happened in a time of the past that is irrelevant to the events being monitored now may lead to incorrect analysis and conclusions.

Once location references are extracted, they can be verified using outside sources and, by considering their spatial, temporal and thematic relevance, their significance determined. Next, all spatially, temporally and thematically relevant locations can be placed on a map and represent the journey taken. This process will be demonstrated in this thesis using Jules Verne’s *Around the World in Eighty Days*, then conclusions drawn as to its applicability to geospatial intelligence.
Jules Verne
A physical copy of *Jules Verne Four Novels* by Ernest Hilbert was purchased and used as the main data source for this project. The following provides a brief introduction to the author, his works and why they are relevant for this project.

Who was Jules Verne?
Jules Verne was born in Nantes, France in 1828. Verne and his brother explored the docks and riverbanks of this harbor city as young boys, witnessing shipping vessels coming in from and departing to distant lands. These events ignited his imagination and he developed a love of travel and exploration. He also had a passion for books and began writing for fun.

Jules Verne’s narratives were originally published in French and were about ‘extraordinary voyages’. Verne sent his first extraordinary voyage, *Five Weeks in a Balloon*, to a publisher when he was 34-years-old. This would be the first of dozens of tales of adventure produced by Verne. His sixth extraordinary voyage, and perhaps what he is best known for, was *Twenty Thousand Leagues Under the Sea*, originally published in 1870. This story may be considered his masterpiece, but his most popular novel, his eleventh extraordinary voyage, originally published in 1873, was *Around the World in Eighty Days*.

Jules Verne died at his home in Amiens, France in 1905. (Verne & Hilbert, 2012).

Relevant characteristics of Verne’s stories
Whether trying to locate the source of the Nile river, an under sea adventure or trying to win a bet by travelling the world on a deadline, all of these stories are about a
journey. As with any journey, there is a starting point and a final destination with many adventures in between. In order to depict these journeys for his readers, Verne used many location references throughout his stories. These references were presented in multiple formats, as both direct and indirect references. The type of location references are important, particularly to use in geospatial intelligence. Direct references are easier to validate than indirect references. Accuracy and inaccuracy of locations, even by a seemingly small distance, can make a big difference in analysis results and, therefore, decisions made based on the results. Direct and indirect references will be discussed further in the Literature Review, Data & Methodology and Results & Discussion sections of this paper.

Fairly unique in fiction for entertainment, Verne includes measures of latitude and longitude. For example, in chapter 18 of Twenty Thousand Leagues Under the Sea, Professor Aronnax, the story’s narrator, relates, “We were close to Vanikoro,…situated in 16° 4’ south latitude, and 164° 32’ east longitude.” Though not an exact location, it is quite easy to find the spot on a map. It can be made more exact if one or more of the other types of location references to be discussed accompany it.

Verne even includes specific address in his stories. These can be the most exact and easiest to verify location references used by the author. For instance, the first sentence of Five Weeks in a Balloon reads, “There was a large audience assembled on the 14th of January, 1862, at the session of the Royal Geographic Society, No. 3 Waterloo Place, London.” Knowing that the story begins in England simplifies the verification of this place.
Not all of the location references in Verne’s stories are as explicit as these two. There can be a lot of ambiguity. For example, chapter 11 of *Around the World in Eighty Days* says, “…they soon came in sight of the great Andaman, the principal of the islands in the Bay of Bengal…” ‘Within sight’ does not give a clear indication of how far the characters were from the island. They could’ve been far enough away that the island was just a spec on the horizon, close enough to see people on shore, or anywhere in-between.

It is because of these and other kinds of location references of real world places and the fact that they can be verified in the same way the USIC or LE may use to verify location reference that appear during the course of their investigations that the stories of Jules Verne will be exploited for this project.

In addition to the location references, there are several more reasons Verne’s stories can be equated to real life intelligence. When text is translated from one language to another, not every word has a counterpart, which may lead to confusion and lost or emphasized meaning. Due to the fact that Verne was born and raised in France, he tends to use syntax and terminology that are used more by the French than residents of other countries. Finally, because Verne was alive mainly in the 19th century, his language and location references line up to that time; not the 21st century lived in now. Meanings and geographic features change over time, sometimes very quickly, so from the release of his first story in 1862 until now, a lot may have changed. The temporal displacement in placenames and situation does have a useful purpose. The USIC and US Department of Defense (DoD) builds resources, including gazetteers from both historical and contemporary sources, and many valuable sources of background information (especially
text-based material) comes from older sources. Bekisz (2015) explores these inconsistencies in placenames over time, and forms a good argument for the need to study this problem, which will ultimately benefit the geointelligence community.
**Thesis Organization**

This paper will be organized into four additional sections: 1) Conceptual Framework, including a literature review and statement of the scientific research questions explored; 2) Data & Methodology, including an overview of the data sources to be used, a description of the data to be collected, and how it will be stored and processed; 3) Results & Discussion; and 4) Conclusions & Future Research.
CONCEPTUAL FRAMEWORK

In the recent past, several efforts have been put forth in the subjects of geographic information retrieval and geoparsing and georeferencing unstructured text in fields such as literary mapping, social media and online documents such as web pages and news websites. However, to the best knowledge of this author, there has been little research done in the application of these procedures in the realm of geospatial intelligence. This section will provide an overview of the published research on these topics and a declaration of the research questions that will be addressed in this thesis. Bekisz’s thesis (2015) presents a study of georeferencing placenames in foreign languages from historical military records, but his work is based on largely structured data accompanied by maps. The work in this thesis uses similar methodology but does not rely on the structure of military records and accompanying maps.

Literature Review

Quoted by many who have done research in literary mapping is Franco Moretti (1999), who says, “A map…is a connection made visible…Of maps, I mean, not as metaphors, and even less as ornaments of discourse, but as analytical tools: that dissect the text in an unusual way, bringing to light relations that would otherwise remain hidden” (p. 3). This is a characterization that can be directly related to work done by the USIC or LE. For example, making associations between events or connecting multiple
persons of interest by places they have visited or paths they have taken. These connections are made more easily when they can be seen. A commonality that may have taken much longer to find otherwise may be found.

Information can be communicated better by the presence of an “image representation of texts” and “geospatial visualization of information…helps users to understand the relationship between data and geographic location” (Abascal-Mena & Lopez-Ornelas, 2010, 829). This is true for works of fiction, so the reader can play out the story within their own mind, as well as real world investigations. When the text does not include pictures or maps, it is up to the writer to provide enough detail to enable the reader to form an accurate picture. It is up to the reader to take the provided information and do something with it. This action can be categorized as literary mapping. Cooper and Gregory (2011) divide literary mapping into two ‘principal categories’: writerly mapping and readerly mapping.

Writerly mapping is defined as “the ways in which an author explicitly explores the relationship between cartography and textuality” (Cooper & Gregory, 2010, 91). Writerly mapping refers to the role actual maps played in the development of the writing of the story and the way the writer “deliberately incorporates cartographical representations of space…within [their] published work” (p. 91-2). In the context of the USIC and LE, writerly mapping would apply to the person of interest being pursued. It would also apply to any member of the USIC or LE writing a report, or other communication, that includes geographic information.
Readerly mapping is defined as “the ways in which an individual may recalibrate this relationship between textual and cartographical representations of geographical space through the reading process” (Cooper & Gregory, 2010, 91). Readerly mapping refers to the way the reader relates what they are reading to the geography it represents. For example, the reader may chose to have an atlas on hand while reading or access a program like Google Earth to visualize the scene or confirm the location of the scene in the story is accurate. They take some kind of action to help form the scene in their mind. The reader may even choose to create their own map based on the text (p. 92), as will be done for this project. Readerly mapping applies to the person(s) in the USIC or LE pursuing persons of interest as well as the intended recipient of the text. An intended recipient can be a second person of interest or associate, or, in the case of a second investigator reading a report, the second investigator would also fall into this category.

The category of writerly mapping in literary mapping and real-world investigations are different. The main difference between the fictional world and the real world is, when the ‘writer’ is a person of interest or suspect, the writer is not usually aware of their role as the writer. The writer does not deliberately include text to explicitly state anything for the reader or to help the reader form a realistic picture of the events. If anything, they would try to be vague or misleading and provide as little information as possible, just in case they were being tracked.

The role of the reader remains essentially the same for fictional or non-fictional text. In the real world, the role for the recipient, their pursuer or a third party remains the same as previously described.
Location references within text can be both direct and indirect. A direct reference could be a specific address, a latitude and longitude or a ‘textimmanent name’ (Reuschel, Piatti & HURNI, 2009, p. 3). A textimmanent name is a name that is explicitly stated within the text. An indirect reference is “deduced indirectly from other sources or researches” such as the context of the surrounding text (ibid).

Though direct references are simpler to locate on a map, they are still subject to ‘synonomy’ (Bordogna, Ghisalberti & Psaila, 2012, p. 107) where multiple variables share a name. There are a few ways this can happen, as discussed in geographic information science literature, i.e., Longley et al. (2015, 154-157) where they are referred to as ambiguities. Geo/Geo ambiguity occurs when one placename refers to multiple places (Amitay, Har'El, Sivan & Soffer, 2004, 273). For instance, there is a city called Hollywood in both California and Florida. In these instances, more information, such as context and additional, temporally relevant, references, is needed to determine which is correct. If a mention of visiting the Chinese Theatre in twenty minutes is present, because of the proximity of the theatre to the two options (Andogah, Bouma & Nerbonne, 2012, 4), it is logical that Hollywood refers to Hollywood, CA and not Hollywood, FL. Geo/Non-Geo ambiguity occurs when placenames have another, non-geographic meaning (Amitay, 2004, 273). A location may share a name with a person (Nesi, Pantaleo, & Tenti, 2014), i.e. Washington. The name could refer to President George Washington, Washington state, Washington DC, etc. The same is true for historical monuments and organizations.
A limitation to indirect references is the “context dependent spatial condition from a location” (Bordogna et al., 2012, 106) such as ‘down the road’ or ‘near the [location]’. The biggest problem with this form of reference is that “people perceive distance in relation to the effort over time spent in travelling from a point to another” (Bordogna et al., 2012, 106). Multiple factors, such as mode of travel, physical condition of the traveller, and more influence this perception. A trip from the east coast to the west coast of the United States is a short trip by plane but a long trip by car and a mile is a much shorter distance for a young athlete than an elderly person with a walker.

Margaret Remillard is quoted by Black, MacDonald and Black (1998, 13) as saying, “Any subject that utilizes spatial information can benefit from [geographic information systems]” when justifying the use of GIS for “historical studies of print culture” (p. 27). The same argument can be made for geospatial intelligence.

Along the lines of Moretti’s view of maps, Black et al. (1998, 13) say, “quantitative data can serve as significant reference points from which qualitative data can be understood”. First, however, the qualitative data must be transformed into quantitative data. A process similar to that used by previous research will be used for this project: 1) Location references will be extracted from the text (manually, in this case), 2) coordinates for the location references will be found using multiple sources, and 3) the coordinates will be placed on a map (Abscal-Mena & Lopez-Ornelas, 2010, 830-31; Cooper & Gregory, 2011, 93; Reuschel et al., 2009, 2; Nesi et al. 2014). This will be detailed further in the ‘Data & Methodology’ section of this paper.
The use of location references to track geocrowdsourced accessibility content is reviewed extensively by Rice et al. (2011, 2012a, 2012b, 2013a, 2013b, 2014, 2015) and Qin et al. (2015a, 2015b) who encounter some of the same difficulties in ambiguity, temporal name shifts, and language translation issues.

**Statement of Scientific Research Questions**

There are many questions that could be explored on this complex, multi-faceted topic. However, not all of them will be explored here. The main questions this project set out to answer are *Can the path taken by an individual in a Jules Verne novel be reliably reconstructed based on unstructured text?, How can ambiguities and uncertainties be resolved and/or clarified?, and How will the georeferencing results from this thesis inform the similar georeferencing activities in the intelligence community?*
DATA & METHODOLOGY

This section will provide detail about the data sources that were be used, how the data was be collected and stored, and how it was used.

Data Sources

The main source of data for this project is *Around the World in eighty Days* by Jules Verne. The version used was published in 2012 in *Jules Verne Four Novels*, which is a compilation of four of the author’s most well known stories, *Five Weeks in a Balloon, A Journey to the Center of the Earth, Twenty Thousand Leagues Under the Sea*, and *Around the World in Eighty Days*, with an introduction to Verne and each of these four novels written by Ernest Hilbert.

Once the location references were obtained (this process will be explained in more detail below), they were georeferenced to their real world locations. This was done with the aid of three main resources, including the National Geospatial-Intelligence Agency’s (NGA) GEONet Names Server (GNS), the Getty Thesaurus of Geographic Names Online (TGN) and Google Earth (GE).

NGA’s GNS is:

The official repository of standard spellings of all foreign geographic names, sanctioned by the United States Board on Geographic Names (US BGN). The database also contains variant spellings (cross-references),
which are useful for finding purposes, as well as non-Roman script spellings of many of these names. All the geographic features in the database contain information about location, administrative division, and quality. (Geonames.nga.mil [GNS], 2015)

NGA’s GNS can be found here: http://geonames.nga.mil/gns/html/index.html.

Getty’s TGN is not a GIS (Geographic Information System) but does provide approximate geographic coordinates for the records it contains. It “is intended to aid cataloging, research, and discovery of art historical, archaeological, and other scholarly information” ("Getty Thesaurus of Geographic Names (Getty Research Institute)", 2016) and covers a broad range of time, from prehistory to present day. These historic records are particularly useful when researching locations during a different period of time than the recent past or present day, such is the case in Around the World in Eighty Days, which takes place in 1872. Getty’s TGN can be found here:

http://www.getty.edu/research/tools/vocabularies/tgn/.

Google Earth is a free downloadable computer program from Google that is very user friendly. It enables the user to add, label and store points and more on a digital globe very quickly. It is a great program for today’s casual user who wants to map a story to employ. Google Earth was downloaded from https://www.google.com/earth/.

These resources were used to verify the name and spelling of the location reference as well as retrieve the corresponding latitude and longitude. If the location reference were not found via these three resources, further research was done.
The *Imperial Gazetteer of India* provided by the Digital South Asia Library (DSAL) was used for locations in India. The DSAL is a “program of the University of Chicago and the Center for Research Libraries” that “provides digital materials for reference and research on South Asia to scholars, public officials, business leaders, and other users,” ("The Digital South Asia Library", 2016). The *Imperial Gazetteer of India* is a historical reference work, originally published in 1881. It covers “the geography, history, economics, and administration of India… listing places' names and giving statistics and summary information…and atlas,” ("The Imperial Gazetteer of India", 2016). Additional maps and websites, that, where relevant, will later be discussed further, were also used.

Microsoft Excel was used to store the data (data table description below) and create data charts. ArcGIS was used to visualize the data and, in conjunction with Google Earth, to help locate indirect references. Railway data for countries, including Great Britain, France, Italy, India and the United States was downloaded from DIVA-GIS’s ‘Free Spatial Data’ page: http://diva-gis.org/Data.

**Data Collection and Processing**

The location references were retrieved from the source text manually. While reading through the story, every reference to a location or measurement was highlighted and recorded. This information was stored in an Excel table. Next, an attempt to locate each reference was made. Once all of the data was collected, it was uploaded into ArcGIS for visualization and reconstructing the journey.
Location References Data Table
Each reference was stored in the table with the information contained within Figure 1.

Figure 1: Fields contained within the location references data table

The first three columns of the table contain information of where within Hilbert’s book the location reference was found. This includes the chapter number, the page number and the paragraph number on that page. The fields are labeled chap_num, page_num and par_num, respectively. The chapter number will be consistent throughout any version of *Around the World in Eighty Days* while the page and paragraph numbers will be specific to the Hilbert version of the story that was used for this project.

Next is the individual location reference, loc_ref, with no additional information to provide clarification or context. Information to provide clarification and context for each reference is stored separately (discussed below).
Reference type (ref_type) and ambiguity type (ambiguity_type) follow. As detailed in the earlier literature review, the reference type are direct or indirect. Direct references are short, typically 1-3 words in length. Indirect references are usually lengthier and can be almost a complete sentence themselves. Ambiguity type are geo/geo, geo/non-geo, dual or NA.

To help provide some clarification and/or context for each location reference, the text of the sentence (ref_sentence) in which the reference was found, the character(s) the reference applies to (character), and a few-word summary of the current events of the story (context) are stored. The reference sentence stored is the full text, verbatim, of the sentence from the book, or the full text before or after a semi-colon (which ever section of the sentence that contains the reference) if one is present. The current context of the story and the character(s) to whom the reference relates was used to determine the relevance of the reference to the journey.

Next is the determination of the relevance of each reference. There are four columns for this purpose. The first is the determination of the spatial relevance (sp_rel) of the reference. A reference is only considered to be spatially relevant for this project if the character the reference applies to is physically present at the location at that moment of the story. Therefore, a reference does not become relevant until the character arrives to the location and becomes irrelevant again once he/she leaves. Next is the temporal relevance (temp_rel), which is closely related to the spatial relevance. A location reference was considered temporally relevant if the character(s) are currently at that location. For determination of relevance for the purposes of this project, if a reference is
spatially relevant, it must also be temporally relevant, and visa versa. The next field is thematic relevance (them_rel). A location reference was determined to be thematically relevant if it is related to Fogg’s journey around the world. This may be a future destination for the travellers, a place they have previously visited on the journey or their current location. Therefore, a location can be thematically relevant without being spatially or temporally relevant. Finally is the overall relevance of the location reference (relevance). In order for a location reference to be considered relevant, it must have been spatially, temporally and thematically relevant. Also, each reference was only considered to be relevant the first time it is determined to be relevant. For example, Fogg arrived at the Reform Club, so it became relevant. ‘The Reform Club’ was mentioned multiple times before he left. Every time it was mentioned was recorded as its own record in the location references data table, with a ‘yes’ in each of the related relevance columns except the overall relevance column. This column will contain ‘yes’ for the first record and ‘no’ for the remaining instances until Fogg leaves the club. If he came back, it once again became relevant and all four columns read ‘yes.’ The amounts of time spent at each location are tracked in the columns that follow.

The next five columns record the method of travel and travel time. The characters used multiple forms of transportation throughout the journey. Most were fairly standard, such as trains and steamers, but some were more non-traditional in the present, every day United States of America, such as an elephant. The transportation vehicles taken to arrive to the location reference are stored in the trans_method column. This information records the method used to arrive to the specified location. There is no column that records the
method used to depart the location. This is an intuitive conclusion based on the record of
the next relevant location. Throughout his journey, Fogg kept a detailed journal of his
trips across the globe where he recorded the date and time of arrival to and departure
from each relevant location. These details are often revealed to the reader, however, in
some instances, the most detail given is that they arrived ‘midday’ which would require
some interpretation. These details are stored in the columns labeled ‘arrival_date,’
‘departure_date,’ ‘arrival_time’ and ‘departure_time.’ Dates and times are only recorded
for relevant location references and only include actual dates and times of their arrival
and departure, and thus exclude all of Fogg’s planned or estimated arrival or departure
dates and times. The relevant locations where no temporal information was provided
were left blank.

Spelling and name changes were recorded in the alt_spell column. This includes
other commonly accepted names or spellings of the location reference as well as what the
location is called in the present day, if the name has changed since 1872. Tracking name
changes over time requires a lot of extra data and is not the focus of this thesis, but is an
important factor to consider. This is especially true for agencies and groups, such as the
US State Department or DoD, working overseas where name changes are a common
occurrence.

The next eight columns contain the geocoded latitude and longitude of the direct
location references or an assumed latitude and longitude for the indirect references. The
GNS, TGN and GE each have a latitude column and a longitude column. If a result was
found using the source it was placed into the corresponding source’s column. For indirect
references, based on the direct references and the context of the current moment in the story, a latitude and longitude were estimated and placed into the ‘assumed_lat’ and ‘assumed_lon’ columns. All latitudes and longitudes were stored as decimal degrees and if no result was found using a source, or latitude and longitude did not need to be estimated, the fields were left blank.

The final column in the location references data table is the ‘Notes’ column. This field contains any important thoughts or facts about each location reference that did not fit into any of the other fields in the table. An example of what might be found in one of these fields is why a determination of relevance was made or why an estimation was made. The raw data can be downloaded as a Microsoft Excel file here: geo.gmu.edu/archive/acave1/Around_the_world_in_80_days.xlsx.

Retrieving Location References from the Text

The collection of location references within Around the World in Eighty Days was done manually. Every time a location reference, or potential location reference, was found, it was highlighted. This helped enable the researcher to easily find the references within the text if the full text needed to be revisited.

In addition to highlighting references, each reference, as well as additional pertinent information, was entered into the location references data table. In order to facilitate easy review of the reference within the book, the chapter number, page number and paragraph number in which the reference is contained was recorded. The chapter number is located at the top center of the first page of the chapter in roman numerals (Figure 2a). The page number is located at the bottom center of each page (Figure 2b).
The paragraph number was manually counted. Each indent, even if only one line, was counted as a paragraph (Figure 2c). If a page started with the continuation of the paragraph ending the previous page, it was also counted as a paragraph.
Figure 2: Example of page from source text. a) chapter number; b) page number; c) paragraph count (Hilbert, 2012, p. 689)

For example, in Figure 2, the first highlighted reference is ‘London.’ It was found in chapter 37, on page 689, in paragraph number 2.
The sentence containing the reference was then recorded. In the case of ‘London’ in figure 2, the field contains “The reader will remember that at five minutes past eight in the evening--about five and twenty hours after the arrival of the travellers in London--Passepartout had been sent by his master to engage the services of the Reverend Samuel Wilson in a certain marriage ceremony, which was to take place the next day.” In the case that the sentence containing the reference included a semi-colon, only the text on the part of the sentence with the reference was recorded. Figure 3 shows an example of this using an excerpt from chapter 33, page 677, paragraph 4:

Figure 3: Example of a location reference within a sentence containing a semi-colon

The highlighted location reference is ‘Banks of Newfoundland’ and the entire text of the sentence is “On the 13th they passed the edge of the Banks of Newfoundland, a dangerous locality; during the winter, especially, there are frequent fogs and heavy gales of wind.” However, only “On the 13th they passed the edge of the Banks of Newfoundland, a dangerous locality;” was recorded in the location references table.
Once this information was recorded, it was categorized as a direct or indirect reference. For example, ‘London’ would be a direct reference. This is because ‘London’ is an explicit statement of a location. An example of an indirect reference can be found in chapter 12, page 593, paragraph 2 (Figure 4).

**Figure 4: Example of an indirect reference (Hilbert, 2012, p. 593)**

In this example, the reference within “They had gone twenty-five miles that day, and an equal distance still separated them from the station of Allahabad” is deemed an indirect reference. Though a specific name for their current location is not provided, it can be generally determined by taking the surrounding content into account. This will be discussed in further detail later in this section. Determining the ambiguity type was done while geocoding the location references.

A short description of the current events of the story was recorded in the ‘context’ column. This was done to aid in refreshing the memory of why a decision was made regarding a particular reference when revisiting the location references data table. Extra
notes to do the same that did not fit into any other category were filled into the ‘Notes’ column of the table. ‘Person, not location’ or ‘name of a train, not location,’ for instance.

If there was an indication of time within the text, either date or time of day, that information was recorded in the corresponding fields of the location references data table. Dates were recorded in a MM/DD/YYYY format. Time of day was stored in 24-hr time, i.e. 1:00 am appears 1:00, while 1:00 pm appears 13:00.

Before continuing, the character to which the reference applied was noted. This element was not yet essential knowledge, but became important during the next step: determining the relevance of each location reference.

**Determining Relevance of Location References**

Determining the relevance of each location reference is important when attempting to reconstruct or predict a person’s path of travel. The following will discuss the rational behind the decisions made regarding relevance for this project. For instance, if it is known that an individual is currently in Baltimore, Maryland but they are talking about Aspen, Colorado, certain considerations arise: spatial, temporal and thematic relevance. Is Aspen spatially relevant to this individual? No, they are in Baltimore. Is Aspen temporally relevant to this individual’s movement? Again, no, they are in Baltimore. Is Aspen thematically relevant to this individual? Maybe. This determination is made based on context. Is the individual discussing a future trip to Aspen? If yes, this location is likely to be thematically relevant. If they are speaking of a memory as a child, it is likely not thematically relevant. It may be thematically relevant if it was the past if
the trip appears to be related to the reason the individual is being tracked. This is the most subjective of these three characteristics.

For this project, a location was only deemed spatially relevant if the character it referred to was physically in, or passing, that location when it was mentioned. The same is true for temporal relevance. If they were not in, or passing, the location at that moment in the story, it was deemed to be temporally irrelevant. Therefore, spatial and thematic relevance were directly related. If it was spatially relevant, it was also temporally relevant, and visa versa. For example, referring back to Figure 3, ‘Banks of Newfoundland’ was deemed to be spatially and temporally relevant because he was passing the location at the moment of the story it was mentioned. It would’ve been spatially and temporally irrelevant before that point and become spatially and temporally irrelevant when it was completely passed.

The next step to determining relevance was thematic relevance. Is this reference related to the objective? It was possible for a reference to be thematically relevant yet not spatially or temporally. This occurred when a reference was made to a location the character had already been while on the journey, or a place they were planning to go on their way to the destination. Continuing the example of ‘Banks of Newfoundland’ in Figure 3, had this reference been made before it was passed but a character said ‘We will pass it!’ it would be deemed thematically relevant, yet spatially and temporally irrelevant. The same was true if it were mentioned again in a context of ‘We were there!’.

Finally, the overall relevance of a reference was determined. In general, in order to be relevant, the reference must’ve been deemed spatially, temporally and thematically
relevant. There are a few exceptions to this rule that will be discussed shortly. An example of the general rule can be seen in Figure 4. Based on the text, it is known that Fogg had not yet reached ‘Allahabad’ but it was the destination. At that point it would be spatially and temporally irrelevant, but thematically relevant, and categorized as not relevant overall, at that point. Upon arrival to ‘Allahabad,’ it became spatially and temporally relevant, remained thematically relevant, and then determined to be relevant overall. When the journey continued beyond ‘Allahabad,’ it returned to being spatially and temporally irrelevant, thematically relevant, and categorized as not relevant overall. Each reference was only recorded as overall relevant at the first mention of it being spatially, temporally, and thematically relevant. If it were said again while it remained spatially, temporally, and thematically relevant, it was recorded as irrelevant overall, simply to avoid multiple data points appearing stacked on top of one another once geocoded and visualized.

Throughout most of the journey taken in *Around the World in Eighty Days*, the reader is informed when a stop is made or a location is visited and/or passed by. These locations, at some point, are all deemed to be spatially and temporally while also being thematically relevant, therefore are overall relevant. Until these locations are reached, and once left, they are not relevant. However, there are a few examples in which a location was visited, but not mentioned until it had already been departed. Figure 5 shows an example of this exception.
The last location before arrival in Suez in which the reader is present with Fogg is passing through Sydenham toward Paris. However, between Paris and Suez, at least two stops are made: Turin and Brindisi. The only time the reader is informed of these stops is when Fogg updates his travel journal in Suez. The reader was not present in those moments of the journey, thus the locations were never spatially or temporally relevant. But, because they were stops along the journey being followed, they are thematically relevant and necessary pieces to the puzzle, and are, consequently, relevant overall.

Lastly, references were only deemed relevant overall if they were in reference to Phileas Fogg. There are instances within the story that Fogg’s servant, Jean Passepartout, was on his own, musing about his circumstances, or was even separated from the group and had to make part of the journey around the world on his own. During these times, the reader is provided some information about his current activities as well as memories. The
memories would be irrelevant because they occurred in the past. But, some instances that are spatially, temporally and thematically relevant to him in that moment, are not relevant to Fogg’s journey. For example, in the beginning of the story, Fogg goes to the Reform Club, leaving Passepartout at home. Fogg’s location at the club is relevant: it’s where the bet is made and the journey officially begins. Passepartout’s whereabouts and activities have no impact on Fogg’s journey.

The next step was to geocode each location reference and categorize its ambiguity type.

**Geocoding Location References**

Linda Hill defines georeferencing as “relating information to geographic location,” (Hill, 2006, 1), going even further define informal and formal georeferencing. She defines informal georeferencing as a “a means of referring to locations, which we use in ordinary discourse using placenames,” and formal georeferencing as “representations based on longitude and latitude coordinates and other spatial referencing systems,” (Hill, 2006, 2). Within *Around the World in Eighty Days*, the location references, as they appear in the text, can be considered informally georeferenced. The process of translating the informal georeferences into formal georeferences is called geocoding (Hill, 2006, 55).

Once the NGA GNS homepage was pulled up, ‘GNS Search – Text Based Page’ was selected. The box to the left of ‘Include Historical records was checked (Figure 6a), the dropdown menu to the right of the ‘Name:’ field was changed to ‘contains’ (Figure 6b), a reference from the location references data table was entered into the ‘Name:’ field (Figure 6c) and ‘Search Database’ was selected (Figure 6d). A pop-up window appeared
while the search was conducted, which would show the results (Figure 7). Notice, in Figure 7a, the number of results, and in Figure 7b, that none of the top four results are in England, where, based on story context, it is known this London should be. This information is used later when categorizing ambiguity type.

![Image of GNS search interface](image)

**Figure 6:** GNS search: a) Include Historical Records; b) contains; c) Enter location reference of interest; d) Search Database; e) Additional search requirements (Geonames.nga.mil [GNS], 2015)

When this occurred, more information was added to the search and it was run again. In this case, the ‘Country Name’ ‘United Kingdom’ was selected. Ideally, the new search would return fewer results, as seen in Figure 7c (inset Figure 7). When an appropriate
match was found, the latitude and longitude in decimal degrees (Figure 7d) were copied into the location references data table.

Figure 7: GNS search results: a) number of results; b) List of results; c) New results with more search requirements; d) latitude and longitude; e) alternate name/spelling (Geonames.nga.mil [GNS], 2015)
A similar process was followed when using the Getty’s TGN. From the main search page provided previously, the location reference of interest was entered into the ‘File Name or ID’ field (Figure 8a) and ‘Search’ selected (Figure 8b).

These results are given alphabetically (Figure 9a) and again there may be a high number (Figure 9b). To narrow down the number of results, the ‘Place Type’ and/or ‘Nation’ fields (Figure 8c) can be filled and the search run again. While this may reduce the number of search results (Figure 10b), the results are still listed alphabetically (Figure 10a).
Figure 9: Initial TGN search results: a) Alphabetical order; b) number of results (“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016)

Figure 10: Second set of TGN results: a) Alphabetical order; b) fewer results (“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016)
Once an appropriate match is found and selected (Figure 11a), the latitude and longitude in decimal degrees given (Figure 11b) were copied into the location references data table.

Figure 11: TGN London result: a) Select appropriate result; b) latitude and longitude in decimal degrees; c) alternate names/spellings (“Getty Thesaurus of Geographic Names (Getty Research Institute)”, 2016)
Figures 6e and 10c show the location of alternate names or spellings for that location. These were recorded in the ‘alt_spell’ column of the location references data table.

Upon opening the desktop version of Google Earth, a search field appears in the top left corner of the screen (Figure 12a). This is where the location was entered. As the entry is typed, search suggestions appear underneath this search bar (Figure 12b). ‘Search’ is selected, and Google Earth finds a location and auto zooms to its location.

Figure 12: Google Earth search: a) Enter location; b) Select Search (Google Earth, 2016)
If an obviously incorrect location appears (i.e. London, Ohio instead of London, England and the result should be in England, not America), more information, if available, should be included in the search bar (‘London, England, not just London). When an appropriate location is found, the location was saved to ‘My Places’ by selecting the yellow pushpin button at the top of the screen (Figure 13a) and filling out the applicable fields within the pop-up ‘New Placemark’ box (Figure 13b). The latitude and longitude (Figure 13c) can be copied from here.

Figure 13: Add point in GE: a) Select add new point; b) fill out location information; c) latitude and longitude (Google Earth, 2016)
A different approach was taken when geocoding the indirect references, as well as the direct references that did not return any results using the three main tools.

This process will be explained using the indirect reference from Figure 4. The reader learns that Fogg and the rest have travelled twenty-five miles that day and had twenty-five more to go before reaching Allahabad. They were travelling on an elephant. There are a very large number of locations that are twenty-five and fifty miles from Allahabad and it is not made clear if they were maintaining a straight course throughout or following a winding path. This increases the number of possibilities. To find an area that represents the possibilities, a multi step process was used. Allahabad is a known location that is geocoded easily. Within ArcMap, a multi-ring buffer was created around it, including rings of twenty-five and fifty miles. Once the buffer was complete, it was saved as a Keyhole Markup Language (KML) and opened in GE. A KML is “a file format used to display geographic data in an Earth browser” like GE ("Keyhole Markup Language | Google Developers", 2016). Figure 14 depicts the buffers inside GE. The edge of the purple ring shows a buffer of twenty-five miles and the yellow ring, fifty. The location twenty-five miles out can be anywhere between Allahabad and the edge of this purple ring and the location of fifty from the edge of the yellow ring and Allahabad. Because it is known the travellers are coming from the southwest (heading in a northeasterly direction), the possibilities of locations are narrowed down. When selecting one point to represent these indirect references, a point was added inside Google Earth on the edge of the ring that corresponded to each reference working in an appropriate
trajectory. The latitude and longitude of these selected points were then recorded into the location references table.

The same process was used to locate other similar indirect references using the standard buffer tool within ArcMap as well as the multi-ring buffer.

Figure 14: Multi-ring buffer used to determine indirect references to locations outside of Allahabad (Google Earth, 2016)(ArcGIS 10.3.1 for Desktop, 2016)
Some references were not found at all using the GNS, TGN or GE. Further research was needed. In some instances this was because the locations simply do not exist. If that was determined to be the case, a location was estimated based purely on the location described of other locations that do exist. ‘Kholby,’ coming from Bombay and fifty miles from Allahabad is one example. Some locations not found with GNS, TGN or GE do exist in some variation. Perhaps the name used by Verne was simply misread or misspelled. But extra research and the context of the story can help find them. ‘Milligaum’ is one of these. In this case, a Google search was done and a hand drawn map of the extent of the Great Indian Peninsula Railway in 1870 was found with ‘Malligaum’ included (Figure 15)("Great Indian Peninsula Railway", 2016), in a location that fit along the path described by Verne. Going back to Google Earth, this general location was found, and with the ‘Local Place Names’ layer turned on, a city called ‘Malegaon’ was found (Figure 16). This name was then looked up in the Imperial Gazetteer of India provided by the DSAL and was indeed around in the time this journey as to have taken place. Because the spelling is similar and the location makes sense, GNS, TGN and GE were then used to search for ‘Malegaon.’ These latitude and longitude were recorded in the location references data table. Rice et al. (2011) discuss the need to store alternative naming within gazetteers to assist with this problem.
Figure 15: Finding Milligaum ("Great Indian Peninsula Railway", 2016)
The final step before reconstructing the journey was to determine the ambiguity type for each reference.

While working through the list of locations contained within the location references data table, each reference was given an ambiguity type assignment. There are four categories applied: 1) geo/geo; 2) geo/non-geo; 3) dual; and 4) NA.

As stated in the Literature Review portion of this paper, a geo/geo ambiguity occurs when one location reference shares a name with another geographical location or feature. London, for instance, can be categorized as having geo/geo ambiguity because
there is more than one ‘London’ around the world: London, England and London, Ohio, to name one case. Whenever a GNS, TGN or GE search returned multiple results, the reference was categorized as having a geo/geo ambiguity. If there were no non-geographical features or locations with the name found, the reference was categorized as having only geo/geo ambiguity.

A geo/non-geo ambiguity occurs when the reference shares a name with a non-geographical feature or location; i.e. the name of a person. Revisiting the ‘London’ example, ‘London’ can be, and is, used as a person’s name. For example, London Fletcher, former inside linebacker for the National Football League’s (NFL) St. Louis Rama, Buffalo Bills and Washington Redskins. If there were no other geographical features or locations with the name found the reference was categorized as having only geo/non-geo ambiguity. A reference incorrectly assumed to be a location reference but was, in fact, in the context of the story, a person’s name (i.e. Brigham Young, the name of a University and the name of a historical figure from Utah, and the eponymous source of feature names in Utah), were categorized as having geo/non-geo ambiguity.

In the case of ‘London,’ both geo/geo and geo/non-geo ambiguity labels apply appropriately. This is not an uncommon phenomenon. Therefore, the ‘ambiguity_type’ field for this reference was completed with a third category, created during the course of this project: dual. Another example of this case is ‘Washington’: President George Washington and Washington D.C., Washington D.C. and Washington state.

The final ambiguity category was ‘NA,’ short for ‘Not Applicable.’ For instance, there are occasions that a cardinal direction is used to refer to a location, such as a region
within a larger area (i.e. the southern states of the United States of America are commonly referred to as the South). If this was the case within the story, it was categorized as having a geo/geo ambiguity. There are instances where a direction is part of a placename that does not accurately describe its true location. For example, ‘The University of Western Ontario’ is not in fact located in the western region of Ontario, Canada. As can be seen in Figure 17 below, the university is in fact located in the southeastern portion of Ontario. This sort of instance did not appear in Around the World in Eighty Days but is an important consideration.
There are also instances within the story that cardinal directions are used as just as what would be expected, a description of direction specifically. Cardinal directions are standard and used worldwide. To this author’s best knowledge, there is nowhere in the world the word ‘west’ is used when an American compass would read ‘north,’ unless the user’s compass is broken or the user does not know how to read it properly. There are multiple references within this story that are not in themselves, location references, such
as the direction from which a gust of wind is coming. A gust of wind from the east is coming from the east; there is no uncertainty about that. Therefore, when Fogg or another character makes these kinds of references, they are categorized as having NA ambiguity.
RESULTS & DISCUSSION

When the location references data table was complete, there were a total of 1,337 location references recorded. This is not an all-inclusive list. Although careful attention was paid when constructing the list, because this process was done manually, human error is always a possibility and some references may have been missed or incorrectly interpreted as non-location references. Of the 1,337 entries, there are 572 unique entries. The five most referenced locations were: London, 61; Bombay, 59; Hong Kong, 56; Calcutta, 33; and India, 31.

These locations accounted for 240 of the location references. These results make sense. London is where Phileas Fogg lives, where he made his bet, and where his journey begins and ends. The next three, Bombay, Hong Kong and Calcutta are major stops on his journey. At all three of these locations, he had to arrive by a specific time to change his method of transportation. At Bombay, he arrived via steamer and departed via train. At Hong Kong, he changed steamers, and at Calcutta, he switched from a train to a steamer. Because he could not have full control over these vessels or their departure times, these were very important stops and therefore highly discussed.

Of the 1,337 references, 590 were spatially and temporally relevant and 1055 were thematically relevant. Finally, 202 references were relevant overall. These results are shown as percentages in the Figure 18 below.
These numbers make sense. Not every location referenced will be visited, so it is easy to have predicted, and agree with after the fact, that not every reference would be spatially and temporally relevant. It also makes sense that more references were thematically relevant than spatially and temporally relevant. As shown above, locations were repeated multiple times. But, the characters did not have to physically be in that location every time it was mentioned. If that were the case there would have been no mention of any location before or after arrival to it. If a location could only be mentioned
while standing in it, it would be impossible to plan anything or communicate plans to others. Finally, 202 of the location references were found to be relevant over all. Again, this makes sense. There cannot be more relevant locations than there are references, nor can there be more relevant locations than are all spatially, temporally and thematically relevant locations. Because references were only deemed relevant one time, it is okay that the amount of thematically relevant references exceeds the number of overall relevant references.

Even if a location was not relevant, it was still geocoded. Of the 1,337 location references, 1,151, or 86%, were direct references and 181, or 14%, were indirect references. There were 5 categorized as ‘NA’ because they turned out to be something other than, but mistaken as, a location. ‘Bringham Young’ (the person, not the university) was referenced four times, for example. These results are shown in Figure 19. This means that the vast majority of references were an explicit name of a location (i.e. country, state, continent, etc.). In theory, it should’ve been easy to find 1,151 of the locations.
It was fairly easy to geocode the locations that still exist whose names are the same or similar today as they were in 1872. But, for the locations that no longer exist or are simply not real locations, this was not the case. In addition to these cases, the 181 indirect locations were harder to find. In some cases, a reference was both indirect and a fictional or no longer existing location. ‘Kholby’ and ‘Rothal’ are prime examples. The example of ‘Kholby’ was given earlier when discussing how indirect references were geocoded using buffers and estimation. Once that location was found, it was possible to estimate the location of ‘Rothal.’ The only information given to indicate the location of
'Rothal’ was in chapter 11, page 589, paragraph 3, 10 and 14: “The train stopped…in the midst of a glade some fifteen miles beyond Rothal,” “At the hamlet of Kholby,” “There’s still a matter of fifty miles to be laid from here to Allahabad,” respectively. ‘Rothal,’ though passed first on the journey could not be found until ‘Kholby’ was found. ‘Kholby’ was found, as described earlier, then used as the center point for a new 15 mile buffer to estimate a location for ‘Rothal.’ The same was true for other locations in India, such as “the village of Kallenger” or ‘Pillaji,’ from where the travellers rescued Aouda from death and she joined the group. It is possible this location was invented for this part of the story. Another element of the references that made geocoding difficult was the ambiguities.

The majority of references were categorized as having both geo/geo and geo/non-geo ambiguity (813 of 1,337). This can be seen in Figure 20. 225 were geo/geo only, 46 were geo/non-geo only, and 223 as having no ambiguity. The references with no ambiguity were cardinal directions, distance references and oceans.
Geo/geo ambiguity made it especially difficult to locate the correct place, especially in India. In 1872, there was more than one place called ‘Bundlecund.’ Two of these locations could make sense along the entire path through India, but only one at the point of the story that it is relevant. Unfortunately, GNS and TGN yield zero results for this spelling. When GE is used, ‘Bundelkhand’ is returned but appears too far east to be relevant during the appropriate time in the story. Further research shows that there is another region with the same name that does fit into the context of the story. There were
also occurrences of a different kind of uncertainty that arose. This was a different name being used to reference the same location.

Simple examples of more than one name referring to the same location include Fogg’s house being referred to by the address once then as ‘his house’ or ‘Saville Row’ for the remainder of the story. Not all occurrences were so simple. A good example is ‘Burdivan.’ This reference appears only once, in chapter 10, page 584, paragraph number 4. It is part of the description of the path of the Great Indian Peninsula Railway. GNS and TGN yield zero results. GE returns ‘Burdwan, West Bengal’ (Figure 21). Further research for ‘Burdivan’ returns the same name.

![Figure 21: Burdivan (Google Earth, 2016)](image.png)

This result was saved as a point in GE. Later on, chapter 14, page 603, paragraph 2, a reference is made to ‘Burdwan.’ This just happens to be the location that was found by
GE and provided through outside searches. For either term, the location provided by GE fits into the path through India at the appropriate time. It is for this reason that it was determined that ‘Burdivan’ and ‘Burdwan’ refer to the same place.

Jules Verne did not have the same resources that are available today and had to rely on maps that were accessible to him in 1872. It is not likely that these were updated as frequently or as easy to verify as GNS, TGN or GE. Therefore, there is always the possibility that the map he used to create his stories was incorrect or outdated. There is also the possibility that Verne simply misread what was written on the maps he used.

After all of the location references were geocoded, the table was loaded into ArcMap on top of a layer of countries obtained from United States Geological Survey (USGS). The initial result containing all 1,337 locations is shown by Figure 22. Figure 23 shows only the 202 relevant points.
Figure 22: All 1,337 locations (ArcGIS 10.3.1 for Desktop, 2016) geo.gmu.edu/archive/acave1/April_Cave.zip

Figure 23: 202 relevant points (ArcGIS 10.3.1 for Desktop, 2016) geo.gmu.edu/archive/acave1/April_Cave.zip
Figure 23 looks much less crowded than Figure 22. In both cases, it is possible to see a path emerge, but it is clearer in Figure 23, especially in Europe and India. Because it is hard to see from the full extent of the trip, the portion of the trip through the United States will be used to explain how the actual path was reconstructed.

Figure 24 shows only the relevant points inside the US. Again, it is clear that there is some kind of path with a few obvious outliers. To narrow down the exact path of the journey, more information was needed. Fortunately, the mode of transportation throughout the story is known.

Figure 24: US relevant points (ArcGIS 10.3.1 for Desktop, 2016) geo.gmu.edu/archive/acave1/April_Cave.zip
Referencing the data table, the main vehicle used travelling in the US was a train. The date must also be considered. The railways in the western states were still very new and being constructed in 1872. Because there were far fewer train tracks then than there are now, the path can be estimated even more accurately. Thus a layer of US railways from 1872 was added to the layout, resulting in Figure 25.

![Figure 25: US railways in 1872 (ArcGIS 10.3.1 for Desktop, 2016) geo.gmu.edu/archive/acave1/April_Cave.zip](image)

The path of travel becomes much clearer. Now it is possible to trace a path through the points along the railway. Knowing the method of travel is important.

There is one leg of the US trip not taken on a train. Fogg pays someone to take them via sledge (or sled) to their next train station. They cannot wait for a new train where they are or they will miss their train at the next station. This part of the story is
clearly fictional and included to add drama to the story. This portion of the journey cannot be traced exactly. There are far too many paths that can be taken when it is not required to stick to a specific path, such as a train track. It is assumed that the shortest path would’ve been taken, which is represented as a straight line. The resulting path across the US is shown in Figure 26.
Once the path was laid, the points were removed and the portions of the trip divided up by method of travel. The orange represents the portion of the trip taken via train, blue the sledge, and purple a steamer (Figure 27).

Figure 27: Fogg's journey across the US divided by travel method (ArcGIS 10.3.1 for Desktop, 2016) geo.gmu.edu/archive/acave1/April_Cave.zip

The same was done for the rest of the journey (Figure 28).
The entirety of the journey has been reconstructed, however, simply looking at the path does not provide the full picture. Figure 28 does not provide detail of direction. The direction of travel is not possible to determine based solely on the information in Figures 22 through 28 without being told “he’s going this way!”

One way to determine the path’s trajectory is to plot the points in order based on the available arrival and departure dates and times. This is the most accurate way to determine direction. If it is known that the person of interest was in New York at noon last Saturday, and in London, England for dinner, it is only logical that he got on a plane in New York, flew in an easterly direction, and landed in London, England in time for dinner. He would not make it to London by dinner by flying westward. Second, logical assumptions can be made as to trajectory based on known locations visited. For example,
in *Around the World in Eighty Days*, Fogg plans to pass through Bombay, Calcutta, San Francisco and New York ending his journey in London; Bombay is confirmed to be the first stop. It would not have been logical to assume he went southeast to Bombay, then turned around, go to New York, then San Francisco, then Calcutta and back to London. These logical assumptions can be used to supplement missing date and time data.

The order of events in the case of *Around the World in Eighty Days* is fairly easy to establish because it is delivered to the reader primarily in chronological order. In the event a location reference was presented to the reader out of order, such as Fogg’s passage through Turin and Brindisi (Figure 5), the reader was given sufficient information to find the event’s correct place in the journey. In real life, however, intelligence is not always gathered or available in chronological order. In order to reliably reconstruct unstructured text data into what can be called “spatiotemporal episodes,” the reader must have “the ability to successfully generate feasible partial orders” (Croitoru, 2009, 64). Croitoru’s *Connecting the Dots: Constructing Spatiotemporal Episodes from Events Schemas* (2009) uses the path of storms to develop a process to create spatiotemporal episodes: “as a storm or a cloud of plume passes over a domain, it moves continuously through the domain and does not exhibit an erratic behavior of ‘jumping’ from one location to another” (Croitoru, 2009, 64). People are similarly constrained to laws of time and physics. This knowledge must be used when reconstructing a person’s movements: “the process moved from point $p_a$ to $p_b$ then one might expect that the process will be detected at $p_i$ before it is detected at $p_j$ (Croitoru, 2009, 64). This is a more formal statement of the example given in the previous paragraph of making logical
assumptions. Though the process described by Croitoru is applied to in-situ sensor data, a similar process, using spatial and temporal constraints (64) and processing and path consistency (67), can be applied to unstructured text data.

When determining location and relevance to a location reference that has been extracted from text, the difference between stated and actual location should be considered. A person’s stated location may not be their actual location. For instance, while driving in a car on I-95 N toward New York, a person is speaking with another on the phone. The one not in the car asks “Where are you?” to which the one in the car replies “Baltimore.” They do not, however, take the Baltimore exit, but continue past on I-95 N to New York. In this case, the stated location is Baltimore, but the actual location is unknown because it is constantly changing, based on speed, along I-95. This issue is present in *Around the World in Eighty Days*. As the characters are moving along the train tracks or through the waters in a steamer, their location is constantly changing. Therefore, most of the location references within should be considered stated locations rather than actual locations. An actual location shows present location, while a stated location can be an indication of past, present or future location. More information is needed to determine if a stated location is an actual location.

When viewing the final path, depicted in Figure 28, one might wonder why this path was taken when it clearly is not the shortest path to circumnavigate the globe. This story represents Jules Verne’s ‘mental map’ of what a trip around the world in 1872 would be. A mental map is “a map which represents the perceptions and knowledge a person has of an area” (Rosenberg, 2016). In developing the story, he was constrained by
the available travel methods at the time as well as his knowledge of the places he wrote about. The main methods used to travel long distances quickly across the ground and oceans in the 19th century, in particular, when he wrote *Around the World in Eighty Days*, were train and steamer, which are the main forms used in the story. He was living in a time that the closest thing to air travel in existence was a hot air balloon. The first airplane was not successfully flown until 1903. Thus, he had to create a path that used the available travel methods.

Verne’s knowledge of the places in his stories is important. He himself never made a trip around the world or visited the foreign places in the story, so most of what he wrote was based on available maps, other information and imagination. If there were, and he knew about, a faster way to travel across Asia than by a combination of a train across India and to the south by steamer, perhaps he would have utilized it. A different, shorter path would have changed the story completely. Eighty days would not have seemed like such a challenging time frame and Fogg would not have met Aouda, who he ultimately married.

These same things should be considered when tracking or pursuing a person of interest. An individual in New York City likely uses the subway, cars and walking as their main methods to move around while an individual in a scarcely populated area in rural Montana would not walk or use subways on a regular basis. Similarly, these two would have different perspectives on how far distances, such as a mile, are. Twenty miles to a New Yorker that usually walks may be much farther than to a Montanan whose closest resource to home is at least twenty miles away and drives everywhere. Also,
similar to the ‘Burdivan’/’Burdwan’ issue where two names were used to describe the same location, different words can be used to describe the same method of travel. For example, a subway in New York City is better known in Washington, D.C. as a metro. The individual(s) analyzing text data to track a person of interest need to be aware of these various possibilities. Or, A New Yorker may talk about distances in ‘blocks’ while a Montanan may use miles or even minutes to describe how far a location is.

Although, thanks to the Internet and current technology, it does seem likely it would be easier to make plans in or discuss an unfamiliar location today than in 1872, it is not likely to do so with 100% accuracy without personal familiarity of the area.

**How this process can be used by the USIC and Law Enforcement**

This same process can be applied to reconstruct movement of individuals for use in the USIC and LE (i.e. persons of interest or suspects). *Around the World in Eighty Days* has some particular characteristics that make it especially suitable to relate to use in these areas.

The first is the role of the detective, Fix, and why he has taken part in Fogg’s journey. Fix has set out to follow Fogg until an arrest warrant form a bank robbery, that Fix believes Fogg responsible for, catches up to him and he can make the arrest. Until the reader knows for sure that Fogg is not guilty, Fix can be viewed as reconstructing the path of a suspected criminal in order to catch up to him and make an arrest.

The second is the way the story is divided into chapters. When planning dealings or taking trips with more than one person, the individuals involved communicate back and forth to coordinate. Today, this is usually done via emails, text messages and phone
calls. The story of *Around the World in Eighty Days* can be viewed as though it were a set of emails between the various travellers and/or Phileas Fogg and the Reform Club members back in London, detailing the trip’s progress. Each chapter can be viewed as a separate email and each chapter title as the subject line because each chapter details a separate topic or section of the trip and each title summarizes the chapter’s contents. The dates and times of some known activities of Fogg’s journey can be viewed as times that geographic locations were available from cell phone data.

Within the communications, location references can be extracted and stored in a way similar to the way described here with some adjustments. Instead of chapter, page and paragraph number, the source of the individual references should be stored. This may include the person of interest’s name, the email account or phone number the text came from, the subject of the message (if applicable), and the date and time the message was sent. Any information necessary that can direct another person to the reference within its source should be stored. If knowledge of method of travel can be determined, whether its stated or inferred (as just done in the New York-London, England scenario), this is ideal and should be stored. Knowing *how* a person of interest is getting around helps to narrow down, or in some instances broaden, the possible paths taken.

The person analyzing the text data to track or pursue a person of interest should be cognizant of the concept of a mental map and what it may mean to their objective. This information may help higher the accuracy of their results. If they can determine the person of interest’s familiarity with a location, they can decide how reliable the location references are. The location references made by an individual unfamiliar with an area
may be considered less reliable than those made by one who is very familiar. The knowledge of the person of interest’s preferred method of travel may also provide invaluable information.
CONCLUSIONS & FUTURE WORK

Conclusions
The processes employed in literary mapping are directly applicable to the different formats of text data that could be gathered by the USIC and LE. The emails, text messages, hand written letters and notes, transcribed conversations, etc. create a story and can be rebuilt as the same. Therefore, the process should be transferable.

The process used here, however, does have its flaws and limitations. For example, there are some limitations to using most gazetteers because they typically store locations only as a single point represented as a single latitude and longitude pair when a line or polygon would be more appropriate and informative. As previously stated, the five most referenced locations in *Around the World in Eighty Days* are London, Bombay, Hong Kong, Calcutta and India. These are all names of cities or countries. It is much more useful to represent these features as polygons. The single location provided by gazetteers, usually the center point of what would be a polygon, can be misleading. For example, Fairfax County in Northern Virginia has an approximate population of 1.3 million people and spans approximately 150 square miles. Google Earth provides an approximate location of 38.908547 latitude and -77.240515 longitude. If Fairfax County were to plan for an event, for example, preparing emergency services and law enforcement for a predicted weather emergency, they would need much more information than the single point provided by a gazetteer to prepare appropriately. Deploying the emergency and law
enforcement resources to that single location would leave the majority of the county uncovered and would not serve much purpose. Some of the action in *Around the World in Eighty Days* that occurs within many of the cities and countries referenced could be mapped better and analyzed differently if a more formal approach to representing features were utilized. Line representations for features like the Mississippi River, which spans “2,350 miles from its source at Lake Itasca…to the Gulf of Mexico,” (*Mississippi River Facts - Mississippi National River & Recreation Area (U.S. National Park Service)*, 2016) would also be more practical and informative.

**Future Research**

There are many ways this subject can, and should be researched in the future.

There are other qualitative subjects that can be mapped. An interesting path of research is to map not only physical locations of the characters in the story but non-tangible features, such as the character’s emotions, as well. In the case of *Around the World in Eighty Days*, Fogg is generally a very calm person that is hard for others to read. At multiple points of the journey, the characters face obstacles that seriously threaten their ability to complete the journey on time. Fogg appears unaffected while the others seem worried and question why he isn’t, or how he can remain so calm. Mapping then comparing each character’s emotions through the duration of the trip is an intriguing concept. How do different people, while experiencing the same events at the same time, feel about what is going on?

Most people do not refer to locations by using specific latitude and longitude or an exact address in their every day conversations. As is the case in *Around the World in*
Eighty Days, general areas, such as a city name, are used. Again, these types of features are best represented as polygons or polylines as opposed to a single point. A way to represent and store polygon and polyline data within a gazetteer would be highly useful, not only for USIC and LE purposes, but for others using gazetteers as well.

A way to formally represent geospatial data quality and uncertainty should be established. This topic is not investigated in this thesis and is very complex and extensive. Elements of Spatial Data Quality edited by Stephen C. Guptill and Joel L. Morrison explores areas that affect uncertainty, such as positional accuracy, logical consistency and temporal information.

A comparison should be done between different translations and languages of the same text. This would require individuals that are fluent in multiple languages to apply the process from start to finish to the same set of text and compare the results. Around the World in Eighty Days is a good source for this because it has been translated into many languages. The results can be compared to view any differences that may emerge.

Another path of research is to obtain actual, publicly available, real-life text data to analyze. This would be more difficult to do as relates to the USIC because so much of that data is classified. However, it can be done for LE if the appropriate unstructured data from a publicized criminal case with an accepted series of events can be obtained. Two interesting case examples are the 2002 DC sniper attacks or the series of events that transpired on June 12, 1994 in Brentwood, CA when O.J. Simpson is believed to have murdered his ex-wife, Nicole Brown Simpson, and her friend, Ronald Goldman. The events of both of these cases can be presumably reconstructed using newspaper articles,
depositions, witness statements, court transcripts and more. Both cases occurred in generally accepted series of events. The results of reconstructing them based off of the unstructured text only can be compared to these generally accepted versions.

Once a standard, practical, and effective process is established, it would be advantageous to find a way to reliably automate the process. There has been research done in automating GIR in unstructured text, but none that are subject to small enough error to be used in the USIC or LE. The ramifications of mistakes are much higher in these areas than mistakes in GIR from twitter, travel narratives or functional stories.

These are just a few of the possible paths of research that can be pursued in the application to geospatial intelligence of reconstructing paths of travel using location references retrieved from unstructured text data.
REFERENCES


ArcGIS 10.3.1 for Desktop (Version 10.3.1.4959). (2016).


*Google Earth (Version 7.1.5.1557).* (2016).


April Cave graduated from Westfield High School, Chantilly, Virginia, in 2006. She received her Bachelor of Science from George Mason University in Criminology, Law & Society with a minor in Intelligence Analysis in 2012.