Forensic Engineering Science: Developing tools for human identification

By

Shaun Fernando & Harrison Boyce

ENG 4001 Research Project

Final report submitted for the degree of

Bachelor of Engineering (Honours)

in

Electrical and Electronic Engineering

Faculty of Science, Engineering and Technology

The University of Adelaide

Project ID 2022s1-EEE-UG-13223

Supervisor: Professor Derek Abbott

Collaborator: Colleen Fitzpatrick

Date submitted: 24/10/2022

Pages in main-matter content: 31

1. Executive Summary

When a person is unable to be identified through traditional means such as direct DNA matching or fingerprinting, newer fields of Forensic Engineering have been used to identify them. One such field is genealogy. Genealogy is the process of analysing a persons genes to find other persons who have matches in their DNA, thus being identifiable as relatives or ancestors of the unidentified individual. By finding many of these relatives a genetic history can be built for the subject and investigations of that history has led to many successful identifications. High profile cases solved using genealogy include discovering the identities of the Golden State Killer, the Canal Killer and more recently, the Somerton Man. This final case was solved during the duration of this project by the supervisor and collaborator of this project, Professor Derek Abbott and Dr Colleen Fitzpatrick. When matches are found for relatives, the basic information of that relative's life is stored in Genealogical Data Communication (GEDCOM) files. This includes many details but the most useful are the location and dates of the relatives birth and death. These dates and locations can be greatly useful in not only tracking potential locations for the subject's birth but also to find additional possible relatives to further fill out their genetic history. However, due to the nature of GEDCOM files being designed only for transfer of information between computers, analysis of this information is incredibly difficult. As such Dr Fitzpatrick has requested a program be made capable of taking the information present in a set of GEDCOM files and creating visual maps of the birth and death locations of each individual present in the files. This project was formed to create the required program and was successful. Written in Python and utilising the Google Geocoding API and a third party program, EarthPoint, the program takes in up to 36 GEDCOM files and converts all the relevant information into Comma Separated Value (.CSV) files which are then converted to Keyhole Markup Language (KML) files for display in Google Earth. For the information inside a GEDCOM file to be considered of use it must include the name of the individual, a date of either the persons birth or death and a location, specific to the level of a city of the same event (birth or death). The program created achieved three main goals as required by Dr Fitzpatrick. First, the program successfully reduced the time it takes to visualse the information in GEDCOM files from a matter of days to a matter of minutes. Second, the program improved on the accuracy of the previous methods, accepting 15% less entries that did not fit the requirements for acceptance and accepting 27.1% more of the correct entries. While the project was a success and meets the requirements laid out by Dr Fitzpatrick improvements to the system are definitely possible. Several alternative APIs were investigated that could potentially increase system accuracy but would incur greater operating costs. Additionally the system could utlisise additional EarthPoint features to animate the markers shown on the Google Earth map to show individual's migration when both birth and death locations are known. Each of these changes would be beneficial in some ways and detrimental in others and for this reason the project team recommends creating these as alternatives to the current system and giving the end user the option to select their preference.

1. Table of Contents

2. Abbreviations

DNA - Deoxyribonucleic acid
SNP- Single nucleotide polymorphism
KML- Keyhole Markup Language
GEDCOM- Genealogical Data Communication
CSV- Comma-Seperated Values
STR- Short Tandem Repeat
CODIS- Combined DNA Index System
CBD- Common By Descent
AL- Ambiguous Locations
NL- Non-Specific Locations
IE- Invalid Entries
VIM- Valid Individuals Missed

3. Key Words

DNA, GEDCOM, KML

4. List of Figures

5. List of Tables

1. Introduction

The following section gives an introduction to the topic of Forensic engineering science and discusses about the aim, motivation, objectives and the overview of the project.

1.1 Background

Human identification has been present for a prolonged period of time. Presently, some of the biometric technologies such as fingerprint, DNA sequence matching, face recognition ,iris identification and retina identification are used to identify humans [3]. The main identification methods used in the present are fingerprint analysis, dental analysis and DNA analysis [4]. However, finding distant relatives of victims is hard with the use of the current methods used to identify humans [1]. Furthermore, techniques used in the past such as visual identification take a large amount of time compared to the rest of the methods [2]. Hence, the use of a software to automate the data visualisation process would reduce the time spent to do the process manually. Furthermore, the use of location markers aids in determining where family branches which the user needs cluster geographically. Similarly, the use of Google Earth simplifies the process as it has the capability of having a timeline of when the family branches clustered during different years. This would aid in simplifying the process of human identification.

Genetic genealogy which is the use of DNA testing in combination with genealogical research documents and historical documents [6]. Genetic genealogical databases have been used in the present to identify historical human remains [5] , human trafficking, disaster victims, criminal activities and to identify unknown human individuals [7]. Genetic genealogy is used in these cases due to it having high amounts of information when compared with DNA databases which are currently available [7].

```
0 @F122@ FAM
1 WIFE @I375@
1 HUSB @I376@
1 MARR
2 DATE ABT 1790
1 CHIL @I377@
1 CHIL @I380@
1 CHIL @I381@
1 CHIL @I360@
```

Figure 1.1 The GEDCOM records of Amelia Dranchman where the bold text shows the family she was born into and is listed as a child [9, Fig.2 ,p.45]

The standard filer type for genealogical data is called GEDCOM [9]. These files were first created in 1984 by The Church of Latter-day Saints [8]. Before GEDCOM files were created people used techniques such as rekeying and direct import [8]. The rekeying technique was stopped as databases started increasing as it was a time consuming process and prone to errors where the applicants for genealogy were required to print data from their old application and rekey it onto a new one [8]. Hence, currently the GEDCOM file type is used which consists of a plain text and unique number system for each individual added to the database [9]. The following numbering system and the text format of

GEDCOM files are and prone to errors where the applicants for genealogy were required to print data from their old application and rekey it onto a new one [8]. Hence, currently the GEDCOM file type is used which consists of a plain text and unique number system for each individual added to the database [9]. The following numbering system and the text format of GEDCOM files are further shown clearly in Figure 1.1 where Amanda is a child of a family of three which consists of a husband, wife and three children with unique numbers as shown.

The information that is stored in GEDCOM files can be extracted and converted into a file format known as Keyhole Markup Language(KML) with the use of the locations of the various individuals stored in GEDCOM files. In terms of KML files, they are used to display geographic data in an Earth style map [32]. In the following project the KML files generated with the use of the information in the GEDCOM files are used to produce geographic data on Google Earth. KML files include lines, polygons, images and placemarks. Moreover, KML can also be used to overlay textures, add HTML content such as font styles, expose rich data and create different camera angles [31].

1.2 Motivation

Visual depictions of geographic data have been regarded as a vastly superior medium for analysing genealogic history. The project was motivated by an industry need for a simple, self contained program capable of displaying the relevant geographic data associated with persons listed in a given GEDCOM file as no such programs currently exist.

1.3 Aims and Scope

The project aims to produce a user-friendly program capable of taking in a GEDCOM file and producing an interactive map displaying the locations and information of each individual listed in the GEDCOM file.

1.4 Objectives

The project contains 4 core objectives. The completion of these objectives will act as milestones for the project. The objective (in order of importance) are as follows

1.4.1 - Objective 1: GEDCOM to Comma-Separated Values (CSV) file Parser

The program created must include a means of parsing a GEDCOM file to produce a CSV file with the relevant information. This includes Name, Date, Location, GEDCOM ID and Event Type.

1.4.2 - Objective 2: Get Coordinates from Location

The program created must be capable of converting from written locations (eg. Adelaide, South Australia) to coordinates.

1.4.3 - Objective 3: Conversion to Keyhole Markup Language (KML)

The program must be capable of converting the above CSV file to a KML file for use in Google Earth.

1.4.4 - Display in Google Earth

The program must be capable of displaying an interactive map via a Google Earth type program, the details of which is given in the KML file.

1.5 Overview

This document reviews the finals results obtained by the project team through completing the project. It will be broken down into several components which include the literature review, method, results and the discussion. The literature review section will provide an overview of relevant information that will give a deeper understanding of the project. The method section will provide a summary of how the whole program was created with the use of the subsystems that were created in relation to the relevant objectives. The results will discuss the results obtained with the new system and compare the obtained results with the existing system. Finally, the discussion section will discuss the limitations, future work and achievements in the project. Furthermore, in terms of achievements, the report will emphasise whether all the objectives in the project were achieved or not, in the case where an objective was not met, the reasons for why it was not met will be discussed.

2. Literature Review

The following chapter provides information that is related to the topic of genetic genealogy. As this is a fairly new technology there are not a lot of literature available on it. Hence, the following section consists of related material to the topic of genetics and genealogy and gives a brief overview about the existing technologies that are present. Moreover, provides an overview of the usages of genetic genealogy in the present and the method is used to generate results with the use of genetic genealogy for the mentioned usages.

2.1 Existing Technologies

Currently there are no programs available that are able to complete the full conversion from GEDCOM to a visual representation. However, there do exist products capable of doing individual steps. These existing technologies have been extremely helpful for the project, both as examples to pull information from to create better programs but also as potential options to incorporate into the project itself. The existing technologies that are useful to the project can broadly be separated into three categories, GEDCOM parsers, Geocoding solutions, and Google Earth projects. Each of these categories give information toward the completion of objectives 1 through 3 respectively.

Many GEDCOM parsers exist, for varying languages and to obtain various details. Github alone hosts hundreds of repositories of GEDCOM parsers in various languages and many formats for the parsed information [10]. Many of these parsers do not retain the information required for this project and are thus not suitable for the project. However several projects such as this [10]by joephayes create basic .csv files with the relevant information. A better example of an existing GEDCOM parser is the GEDxlate program from GEDmagic. This program takes a GEDCOM file and translates the data into a variety of formats, including .csv. The program runs quickly and obtains all information required. However this program is still only capable of doing the conversion and thus would not satisfy the aim of the project.

For Geocoding, there are many possible solutions. As with GEDCOM parsers many examples exist on repository sites such as Github but the broadest and most reliable solution found was the Google Geocoding API. Calls to this API convert addresses, both specific and general to geographic coordinates which can then be used to place markers on a map. Beyond the simple geocoding, there are options available for filtering by country, postal code and more, as well as the ability to bias toward certain regions. This API perfectly covers everything required for this objective and with its built in synergy with both Google Earth and Google Maps this is a potential solution for this objective in the project. With a billing cost of 5 USD per 1000 calls the program created would thus have to have some way of financing itself if this existing technology were to be implemented.

For the visualisation of the geographic information, Google Earth was recommended to be used by both the supervisor and collaborator of the project. The program Earth Point has been used by genealogists to display the geographic information obtained from GEDCOM files and is quite robust, being capable of cycling though the location markers as a timeline is moved and separating the markers into related clusters. However, this Google Earth Tool is only functional for the United States and therefore lacks some of the functionality desired from this project. The collaborator for this project, Colleen Fitzpatrick has said she is happy with the Earth Point program and stated she is ok with that being used as a separate program to complete the third objective of the project if needed.

While no programs currently exist capable of meeting all objectives of the project, a program was previously available that could cover the first three objectives, and combined with Earth Point was a viable solution for experts in the field such as Colleen Fitzpatrick. Unfortunately from speaking with Colleen the project team found this program, Gen Detective, was no longer functional in the way it previously had been and was thus no longer viable for her work. As such the project team began further investigations into existing literature in the field to better understand what was required for the project.

2.2 DNA and genealogy

In terms of genetic genealogy, the use of genetic genealogy aids in creating a hierarchy of family or biological relationships between individuals. Furthermore, genetic genealogy can also determine the type of relationship between individuals [11]. Genealogists have the option of testing three different types of information in DNA when looking for genetic connections between individuals which consists of Y-chromosomal DNA, mitochondrial DNA and autosomal DNA. Mitochondrial DNA consists information of both males and females [21].

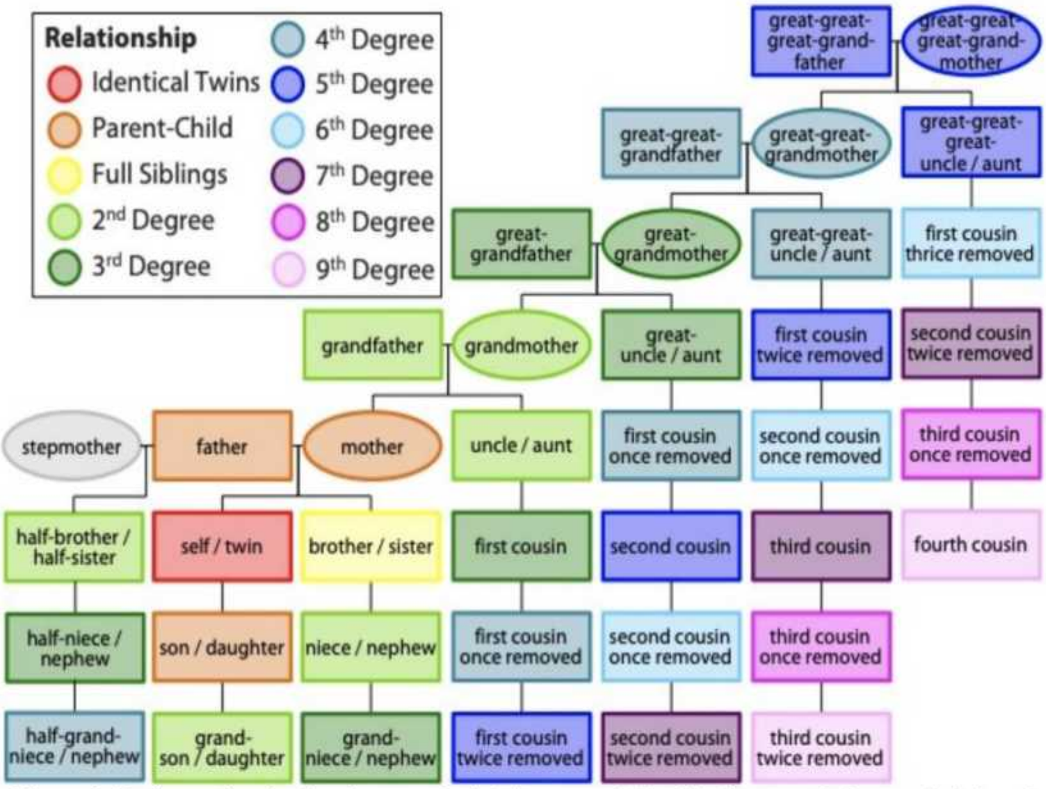


Figure 1: Pedigree showing the degrees of relatedness, as defined by the expected amount of shared DNA. Each relationship is defined with respect to the red "self / twin" box.

Figure 2.1 The degree of relatedness compared with self/identical twin [6, p.4]

A common ancestor who has similar genetic characteristics can be found with the use of mitochondrial DNA and Y-chromosomal DNA. Furthermore, both DNA types can also be used to identify migration [11]. The three types of information in DNA mentioned play a huge role in determining the locations in which an individual's ancestors lived in and the type of relationship the individual has with them. As seen in Figure 2.1, a hierarchy of family and their relation to one another can be produced as such with the use of genetic genealogical databases which is highly beneficial [6]. Hence, in the context of Figure 2.1, it can be seen that the great-great-great grandfather of the identical twin can be displayed as such and the relations between other members can also be displayed.

2.2.1 Y Chromosomal DNA

Y-chromosomal DNA only consist of information on males. Hence, the information in Y chromosomal DNA can be used to determine a male individual's paternal descent [11]. Figure 2.2 shows an overview on how paternal descent can be found with the use of Y chromosomal DNA.

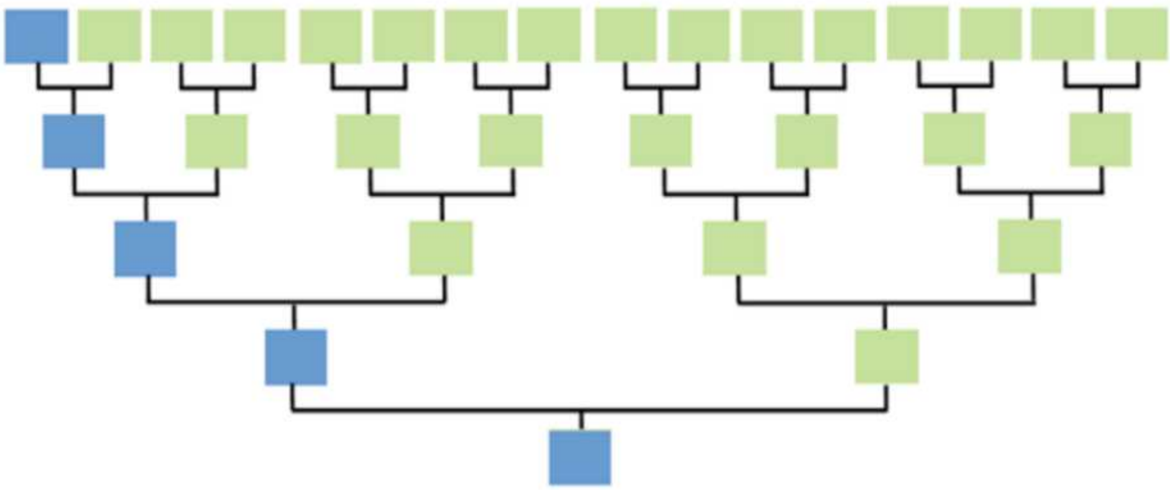


Figure 2.2 Overview Paternal descent which is shown in the colour Blue [17, p.1]

The information in the Y chromosomal DNA can also be used to study the spread of humans who live in the present in respect to humans who lived in the past [12]. All individual genes also consists of alleles which are considered as an alternate form of a given gene [14]. The Y chromosomal T allele to C allele transition are mostly restricted to individuals in Asia and Europe [15]. With the use of the following information in Y chromosomal DNA genealogists and archeologists suggest that parts in East and north Asia have been continuously inhabited during the past 35,000 years [15]. Moreover, the research done by the genealogists show that the use of Y chromosomal DNA aids in comparing the maternal and paternal information which aided in providing evidence of Asian paternal contribution to northern European populations [15].

The study done by T. Zerjal et al. [15] consisted of 1154 males and as seen in Table 2.1 it can be seen that there is a difference in T and C allele amounts in different populations in different countries. Furthermore, as seen in Table 2.1 populations in southern Asian, southern European, American, African and Oceanic only contained the T allele. Both T and C alleles were found in Asian and northern European populations as seen in Table 2.1 [15]. Hence, the study provides evidence which is required to show Y chromosomal DNA aids in finding the distribution of the male population by comparing information from the paternal and maternal information and by comparing the results of the amount of C and T alleles.

Table 2-1 C and T allele distribution in different continents and regions [15, p.1176]

Frequency of T and C Alleles

Continent/Region and Population	No. of Alleles		
	T	C	Total
Africa:			
Kenyan	14	0	14
San	9	0	9
Algerian	27	0	27
Other	9	0	9
Europe:			
Italian	13	0	13
Albanian	10	0	10
Hungarian	39	0	39
Basque	26	0	26
German	71	0	71
United Kingdom	25	0	25
Icelandic	28	0	28
Norwegian	51	2	53
Finn	10	11	21
Saami	9	3	12
Estonian	10	9	19
Mari/Morkinsky	8	8	16
Mari/Gornomariysy	13	7	20
Mari/Orshansky	13	0	13
Mordva	7	2	9
Russian	17	3	20
Other	12	0	12
Asia:			
Indian	53	0	53
Sri Lankan	22	0	22
Buryat	47	64	111
Khalkh	46	1	47
Mjangad	1	1	2
Other Mongolian	14	0	14
Khalimag	0	1	1
Yakut	3	18	21
Altai	28	0	28
Keti	12	0	12
Evenki	25	0	25
Chinese	43	0	43
Japanese	163	1	164
Other	33	0	33
America:			
Amerindian, North	2	0	2
Amerindian, Central	3	0	3
Amerindian, South	22	0	22
Oceania:			
Trobriand Islands	63	0	63
Roro	13	0	13
Other	9	0	9
Total:	1,023	131	1,154
. . .:			
Chimpanzee	4	0	4
. . .:			
Orangutan	2	0	2

2.2.2 Mitochondrial DNA

Mitochondrial DNA is only passed down to an individual by the mother and is present in both males and females [16]. Mitochondrial DNA can be used to determine the maternal descent of individuals as it is only passed down by the mother of an individual [11]. An overview of how maternal descent is found with the use of mitochondrial DNA is found in Figure 2.3 below.

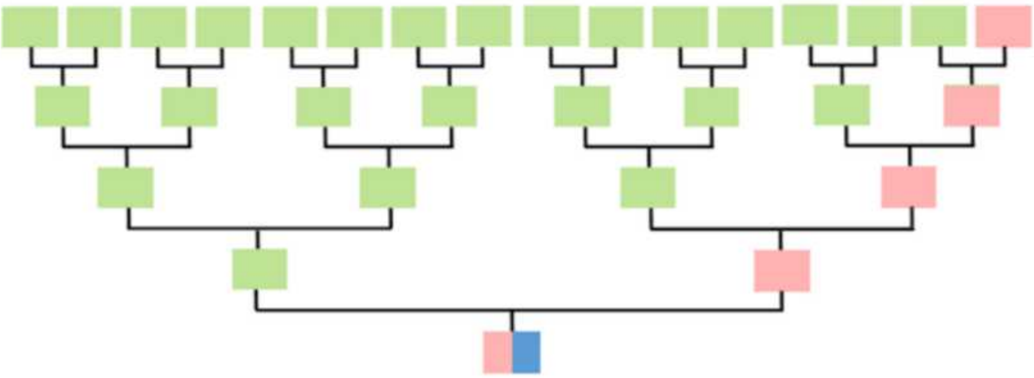


Figure 2.3 Overview of Maternal descent of an individual shown with the colour pink [17, p.1]

Mitochondrial DNA has been used and currently used to determine population genetics studies and human evolutionary studies. One of the main reasons mitochondrial DNA is beneficial is due to the fact that it lacks recombination which is highly beneficial in terms of exploring genealogical relationships with individuals at a regional and continental scale [18]. The lack of recombination in mitochondrial DNA means that the sequence of DNA does not change and remains the same for many generations. Furthermore, mitochondrial DNA mutates much faster when compared with the DNA from the nucleus [18]. A mutation which is highly beneficial in tracing maternal descents is called single nucleotide polymorphism(SNP) which is shown in Figure 2.4 [19]. DNA consists of four chemical bases which store information specific to each individual which are Adenine (A), cytosine(C), thymine(T) and guanine(G) [13].

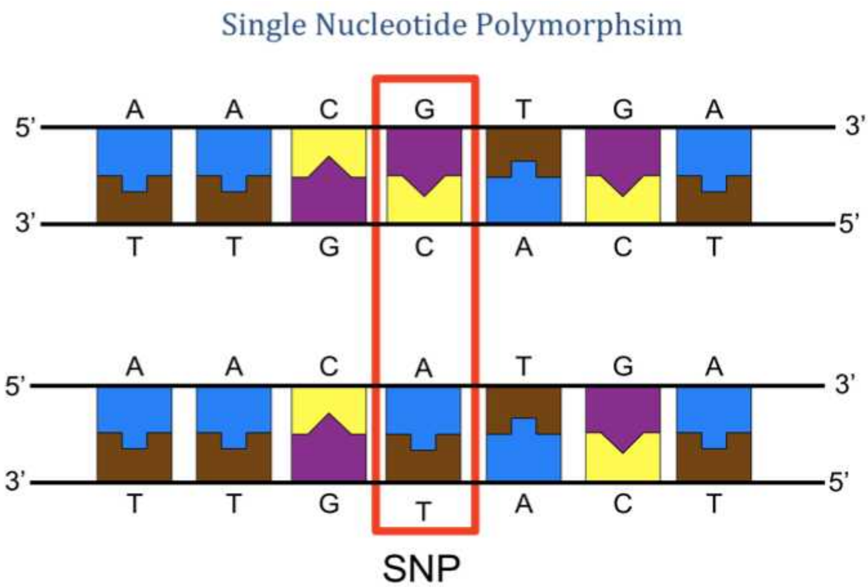


Figure 2.4 SNP mutation in a DNA sequence where in the fourth nucleotide is different in the DNA sequences in the two people [19, p.1]

The single nucleotide polymorphism mutation causes a single base pair in the DNA sequence to be swapped out for a different nucleotide. Hence, in the sense of Figure 2.4, it can be seen that the base pair of guanine and cytosine of person one is different to the base pair of adenine and thymine of person two. Hence, mitochondrial DNA can be used to find maternal descendants by comparing patterns of SNP with other individuals. Individuals who have a similar patterned SNP will be considered as to having maternal lineage [19].

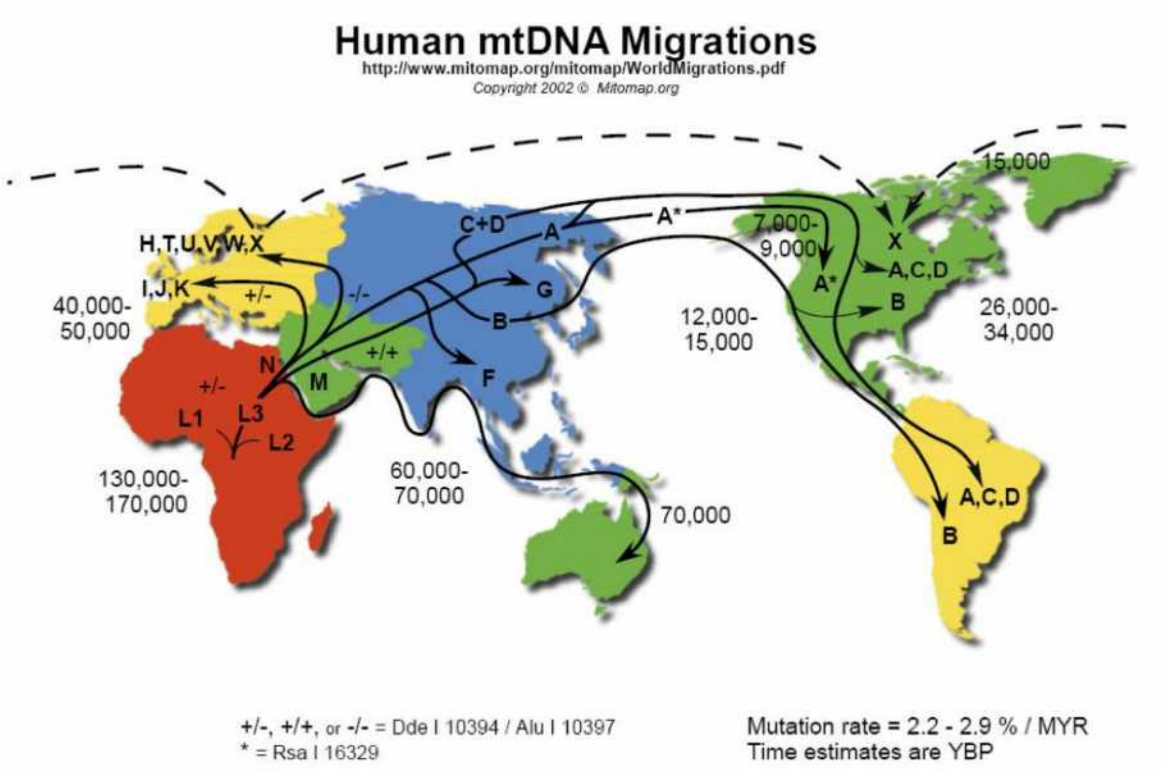


Figure 2.5 The use of mitochondrial DNA to determine haplogroup distribution around the world between 170,000 years ago until 2002 [19, p.2]

Haplogroups are defined as sequences of mitochondrial DNA which have had polymorphism variations which have occurred over more than 150,000 years and correlate to the geographic locations of populations through maternal lineage [20]. When considering Figure 2.5, it can be observed that the most population in the map is shown in Africa and two haplogroups consisting of group M and group N branching out to other parts of the world where the haplogroup M branching out to Asia and haplogroup N branching out to Europe. Furthermore, Figure 2.5 shows that some populations turned in the direction of northern Asia which is shown by haplogroup A in the map. Hence, the use of mitochondrial DNA aids in finding the distribution of populations with the use of different haplogroups with their maternal information.

2.2.3 Autosomal DNA

Autosomal DNA consists of information on both males and females. This means that with the use of autosomal DNA information on both maternal and paternal descents can be obtained [11]. Autosomal DNA provides information on most of an individual’s DNA [21]. The downside of using autosomal DNA is that it is unable to trace ancestors who are beyond approximately 300 years. Hence, autosomal DNA can be used to find ancestors in recent generations. Furthermore, autosomal DNA can be used to identify an ethnicity of an individual by comparing the results with different people in different ethnicities [21]. Presently, the combination of both mitochondrial DNA with autosomal DNA is being used to determine an individual’s ancestors, the spread of population and their ethnicity [22].

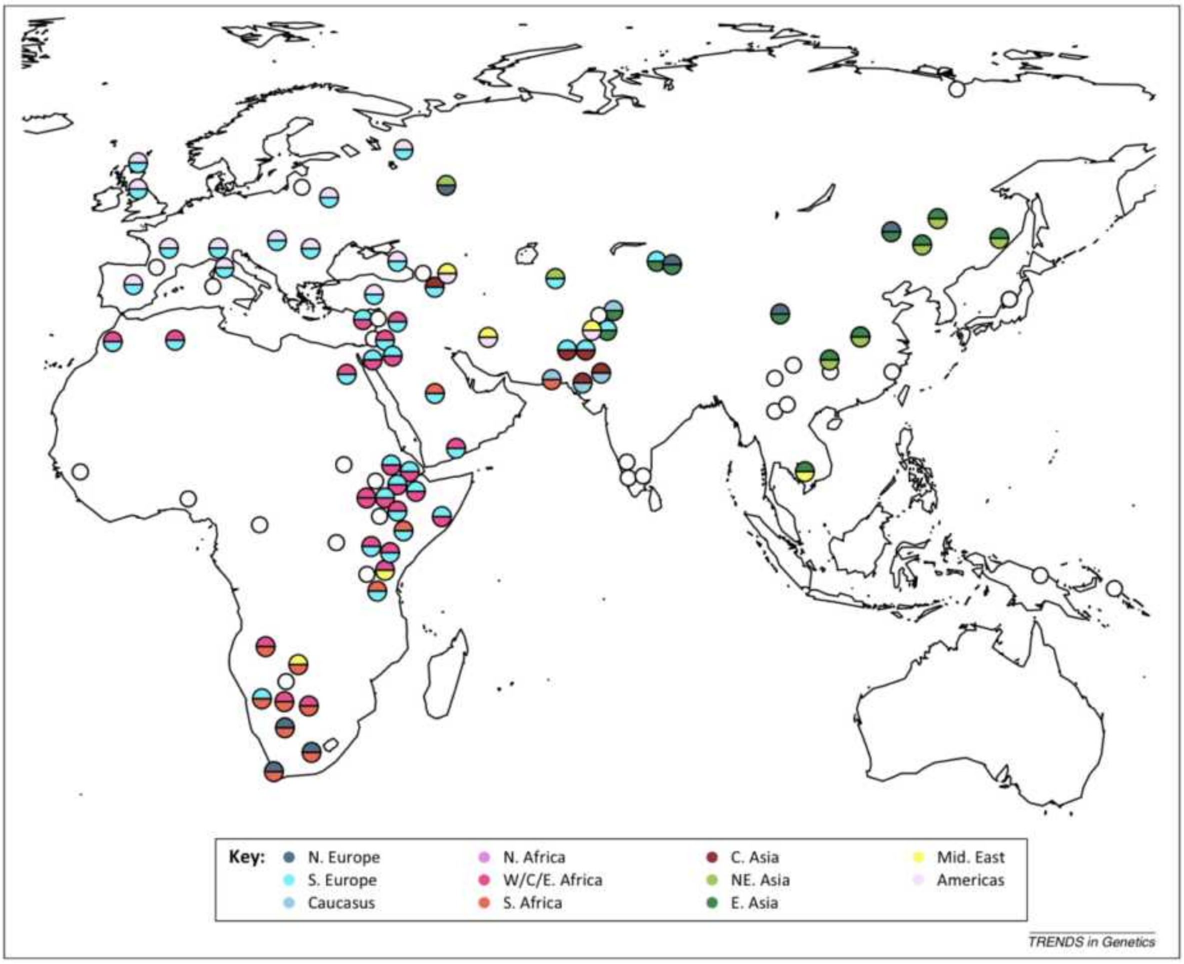


Figure 2.6 Distribution of population where empty circles mean there were no significant evidence on the population[22, p.381]

The use of mitochondrial DNA with autosomal DNA has given results on how the population of people have been distributed among different countries. The distribution of people has been shown in Figure 2.6 where it can be seen that different haplogroups have combined together and developed populations [22]. Hence, autosomal DNA can be used to determine and trace the genealogy of an individual. Furthermore, with the combination of autosomal DNA and mitochondrial DNA the distribution of population of an individual's ancestors can be found.

2.3 Current methods of testing genetic genealogy

The demand for genetic genealogy currently is on the rise in the field of forensic science as this is a fairly new technology. The tests which have been used currently mostly consist of comparison of an individual with a standard STR- based DNA profile [23]. Furthermore, high resolution commercial direct-to-consumer tests analysed under high-density microarrays are also being used presently as they give information of distant relatives compared to standard STR tests [1].

The amount of SNP genotype data has been increasing exponentially as the cost of testing has significantly reduced compared to the past. Similarly, the SNP databases available in the present is high and increasing exponentially [24]. This is due to the fact that many companies such as GEDmatch, deCODE genetics, AncestryDNA, MyHeritage and 23andMe have entered the field of genetic genealogy [25].

Table 2-2 The top ten countries with most frequent GEDmatch uploads [1, p.2]

	Country	Users		Country	Users
1	United States	65%	6	Germany	1%
2	United Kingdom	9%	7	Sweden	1%
3	Canada	6%	8	Ireland	1%
4	Australia	4%	9	New Zealand	1%
5	France	2%	10	Netherlands	1%

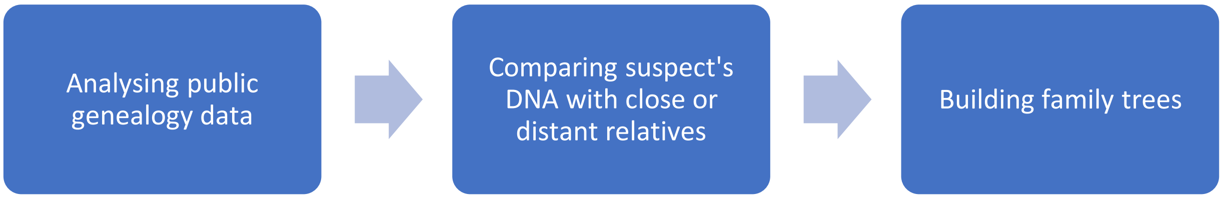
The high amount of SNP databases have aided in the genetic genealogy technology to move forward. As seen in Table 2.2, users in countries such as United States, United Kingdom, Canada and Australia are using companies such as GEDmatch to upload their SNP data which is open to the public which aids in comparing related pairs with the use of the large scale SNP data available currently [1].

2.5 Current usages of genetic genealogy

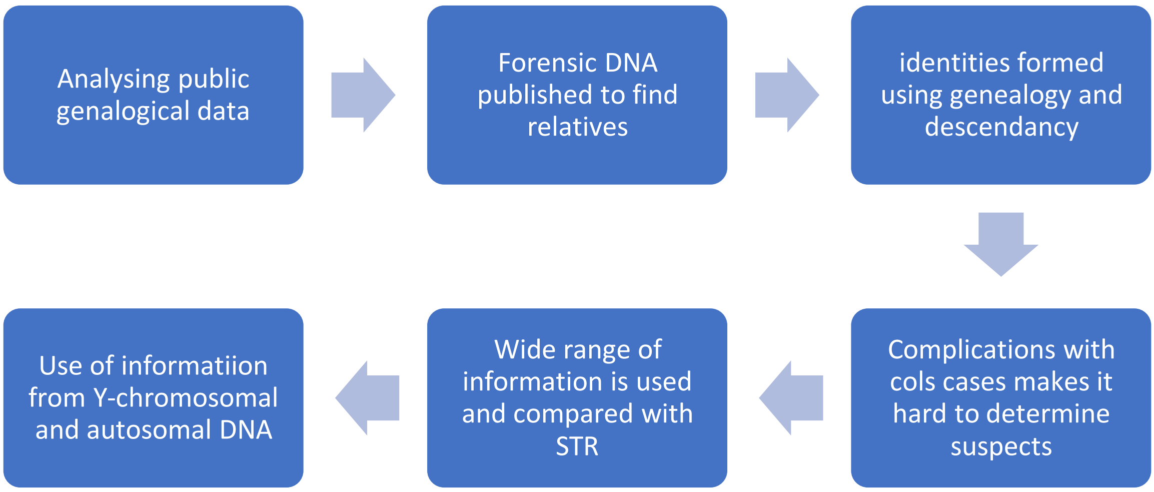
Currently genetic genealogy has been used mostly by law enforcements for their investigations which mainly consists of crimes, cold cases and missing persons [28]. Furthermore, it is also used for finding an individual's ancestry upon request and also to find missing deceased individuals due to circumstances such as war and natural disasters.

2.5.1 Law enforcements usages

Law enforcements in the United States have been able to solve a high number of cases with the use of the publicly available genealogy information. In terms of the method used in solving these investigations which mainly involve crimes, firstly, the forensic genetic genealogy information from databases which consists of genetic private data of citizens who are not under any investigation or have committed crimes are reviewed. Then the crime scene DNA is analysed with the genealogy data which is present by comparing it with close or distant relatives of the suspect and building family trees of all the different connections that can be formed which could be many generations long. A summary of the following process is shown below.



The following method is used by law enforcement to identify potential suspects. This is only performed by law enforcement if the suspect's DNA has not matched with the genetic profiles present in the Combined DNA Index System (CODIS) for a known offender or someone who has been arrested before. Furthermore, the following process is extended for cold case investigations. In the case of cold cases, after the forensic DNA is published to the public to find relatives, a list of all the possible identities with the use of genealogy and descendancy research is generated [28]. As there are many complicating factors with cold cases, the found relatives with the use of the forensic DNA are narrowed down with the use of STR and a wide range of information that is present [26]. In terms of the wide range of information this can include the age of the remains and the part of a family tree which moved to a location which is close to the crime scene. Moreover, in some cases the use of Y-chromosomal DNA and autosomal DNA can be beneficial in finding surnames of the suspects [27]. A simpler version of the following process is shown below in the diagram below.



A table showing a few cases which were solved with the use of forensic genetic genealogy is shown below in table 2.3.

Table 2-3 Cases solved in the US with the use of forensic genetic genealogy [11, p.104]

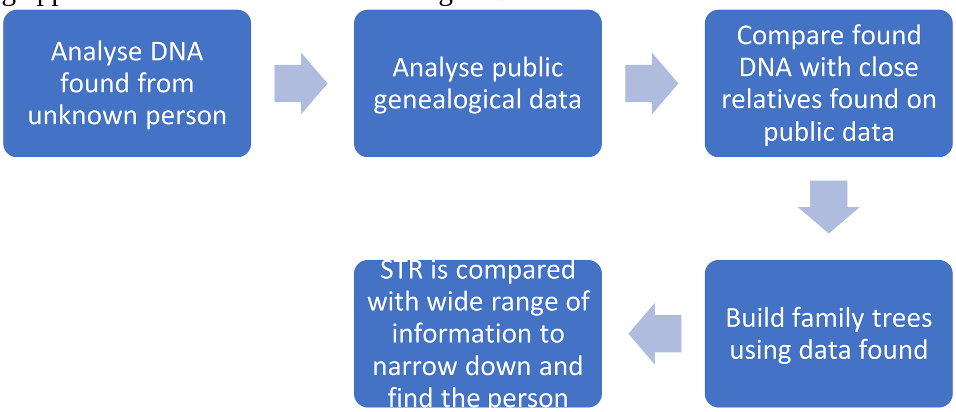
	Location	Case	Year(s)	Identified as	Date announced	Genetic genealogist
1	California	Multiple homicides and sexual Assaults —"Golden State Killer"	1974–1986	Joseph James DeAngelo	April 24, 2018	Barbara Rae-Venter
2	Snohomish County, WA	Double homicide of Jay Cook (20) and Tanya Van Cuylenborg (18)	1987	William Earl Talbott II	May 21, 2018	Parabon
3	Tacoma, WA	Homicide of Michella Welch (12)	1986	Gary Charles Hartman	June 20, 2018	Parabon
4	Lancaster, PA	Homicide of Christy Mirack (25)	1992	Raymond Charles Rowe[b]	June 25, 2018	Parabon
5	Brazos County, TX	Homicide of Virginia Freeman (40)	1980	James Otto Earhart[a]	June 25, 2018	Parabon
6	Fort Wayne, IN	Homicide of April Tinsley (8)	1988	John Dale Miller[b]	July 15, 2018	Parabon
7	Woonsocket, RI	Homicide of constance Gauthier (81)	2016	Matthew Norman Dessault	July 18, 2018	Parabon
8	St. George, UT	Sexual Assault of Carla Brooks (79)	2018	Spencer Glen Monnett[b]	July 28, 2018	Parabon
9	Fayetteville, NC	Multiple Sexual Assaults —"Ramsey Street Rapist"	2006–2008	Darold Wayne Bowden	August 22, 2018	Parabon
10	Champaign County, IL	Homicide of Holly Cassano (22)	2009	Michael F. A. Henslick	August 29, 2018	Parabon
11	Montgomery County, MD	Multiple Sexual Assaults	2007–2011	Marlon Michael Alexander	September 14, 2018	Parabon
12	Sarasota, FL	Homicide of Deborah Dalzell (47)	1999	Luke Edward Fleming	September 19, 2018	Parabon and Barbara Rae-Venter
13	California	Multiple Sexual Assaults —"NorCal Rapist"	1991–2006	Roy Waller	September 21, 2018	Law Enforcement
14	Greenville, SC; Memphis, TN; Portageville, MO	Multiple Homicides and Sexual Assaults	1990–1998	Robert Eugene Brashers[a]	October 5, 2018	Parabon
15	Starkville, MS	Double homicide of Betty Jones (65) and Kathryn Crigler (81)	1990	Michael W. DeVaughn	October 8, 2018	Parabon
16	Greenbrier, AR	Homicide of Pam Felkins (32)	1990	Edward Keith Renegar[a]	October 29, 2018	Parabon
17	Fulton County, GA	Homicide of Lorrie Ann Smith (28)	1997	Jerry Lee	November 1, 2018	Parabon
18	Anne Arundel County, MD	Homicide of Michael Temple (29)	2010	Fred Lee Frampton, Jr.	November 2, 2018	Parabon
19	Orlando, FL	Homicide of Christine Franke (25)	2001	Benjamin L. Holmes	November 5, 2018	Parabon & Florida Dept. of Law Enforcement
20	Carlsbad, CA	Homicide of Jodine Serrin (39)	2007	David Mabrito[a]	November 13, 2018	Parabon and Barbara Rae-Venter
21	Santa Clara, CA	Homicide of Leslie Marie Perlov (21)	1973	John Arthur Getreu	November 21, 2018	Parabon
22	College Station, TX	Multiple Sexual Assaults	2018	Christopher Quinn Williams	December 12, 2018	Parabon
23	Cedar Rapids, IA	Homicide of Michelle Martinko (18)	1979	Jerry Lynn Burns	December 19, 2018	Parabon
24	Hernando County, FL	Sexual Assault of Unnamed Victim (12)	1983	William L. Nichols[a]	January 10, 2019	Parabon
25	Orange County, CA	Sexual Assaults of Two Unnamed victims (9 and 31)	1995 & 1998	Kevin Konther	January 11, 2019	Law Enforcement
26	La Mesa, CA	Homicide of Scott Martinez (47)	2006	Zachary Aaron Bunney	January 24, 2019	Parabon
27	Fremont, CA	Homicide of Jack Upton (30)	1990	Russell Guerrero	January 24, 2019	Parabon
28	Portland, OR	Homicide of Anna Marie Hlavka (20)	1979	Jerry Walter McFadden[a]	January 31, 2019	Parabon

[a] Deceased.
[b] Pled guilty.

As seen in table 2.3, it can be observed that cold cases such as the "Golden State Killer" have been successfully solved with the use of forensic genetic genealogy [11].

2.5.2 Usages in deceased unknown individuals

In terms of the usages of forensic genetic genealogy in deceased unknown individuals, an identical approach to that of the law enforcements is used to find unknown individuals. A summary of the following approach is shown below in the diagram.



The information used to draw out family trees in both cases mostly consists of birth information if they are available, age, death, marriage certificates, grave records and other information that is available publicly [27-29]. The process of drawing out family trees is a lengthy process therefore, forensic scientists tend to limit the relatedness according to the 3rd closest cousin or closer once a best possible match is found with a relative[27].

An example of a recently deceased unknown individual that was managed to be found with the use of genetic genealogy is the Somerton Man and a figure of the following individual is shown below in figure 2.7. The following individual was unknown for over 70 years and was able to be solved to a certain individual with the aid of genetic genealogy.

Figure 2.7 A picture of the Somerton Man [30]

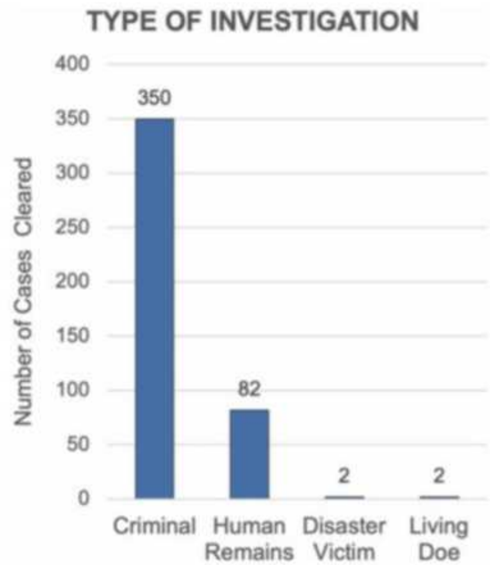
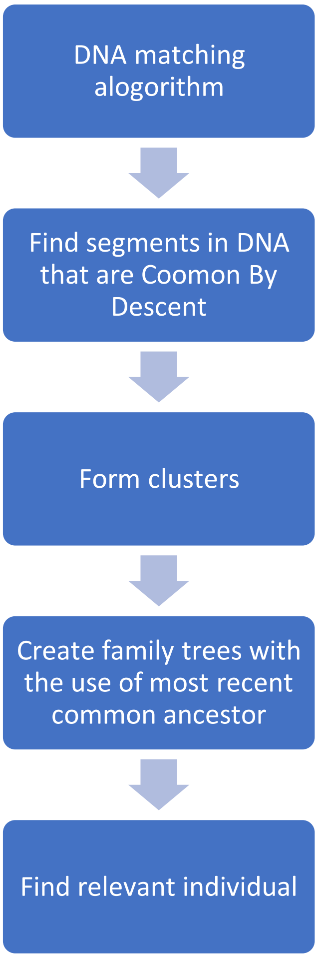


Figure 2.8 Cases that have been solved with the use of forensic genetic genealogy in the USA [26]

When considering the data shown in figure 2.8, it can be observed that this new rising technology of using genetic genealogy to solve various types of cases which includes as shown in this case disaster victim, living does, human remains and criminals is becoming successful.

2.5.3 Use of matches to create family trees

When considering how genealogists find matches to create family trees, these are done with the use of DNA matching algorithms where two individuals who share common segments of DNA are considered a match [26-27]. The following segments of DNA which have matches are passed on to the individuals from a common ancestor and are named as Common By Descent (CBD) segments [29]. Once common segments are identified in various individuals, genetic genealogists can form clusters with use of various shared matches where they can work with genetic networks to identify the most recent common ancestor for each of the clusters that have been found [27]. Genetic genealogists can then form family trees with the use of the most recent common ancestor found and then identify all the common descendants of the most recent common ancestor. Therefore, by performing the following method the person of interest can be found as they would normally lie among the group of persons found with the method of developing family trees with the use of the most common ancestor for each cluster formed [26]. A more simpler and summarised version of the following process has been shown in the diagram below.



The following method has generally been used by genetic genealogists in ways to solve cases which involve crimes, unknown deceased persons and human remains. The creation of family trees with the use of the method mentioned above is a major and important step in finding specific persons that need to be found.

2.6 Literature Summary

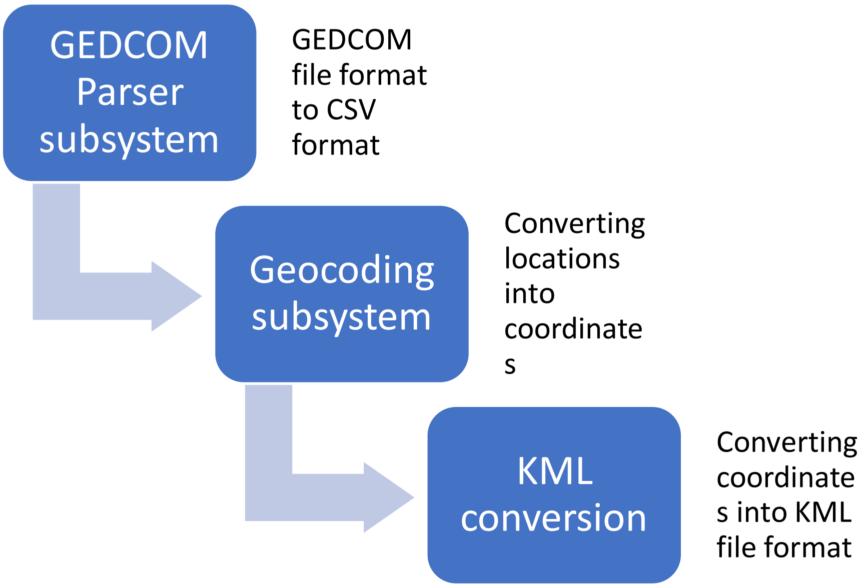
In summary, it can be observed in the following chapter that the types of information found in DNA such as autosomal DNA, mitochondrial DNA and Y chromosomal DNA aid in finding relationships which an individual has with their ancestors, their ethnicity and the movements into different regions of the world in the past generations. These key aspects of DNA aid in determining an individual's background and aid in analysing the information found in the results during the progress of the project. Furthermore, it can also be observed that genetic genealogy is a rising technological field and currently consists of various usages and mainly is an important factor in solving missing persons cases and crimes.

3. Method

The following section gives an overview of the method used to develop the program required for the project. The method section consists of three main sections.

1. GEDCOM parser subsystem
2. Geocoding subsystem
3. KML conversion

The combination of the following sections and the proper functionality of each section provides the functionality for the program to function efficiently and accurately. The key steps taken to develop each section in the program will also be discussed in the following section. Moreover, the challenges faced in terms of developing each section will also be discussed where applicable.

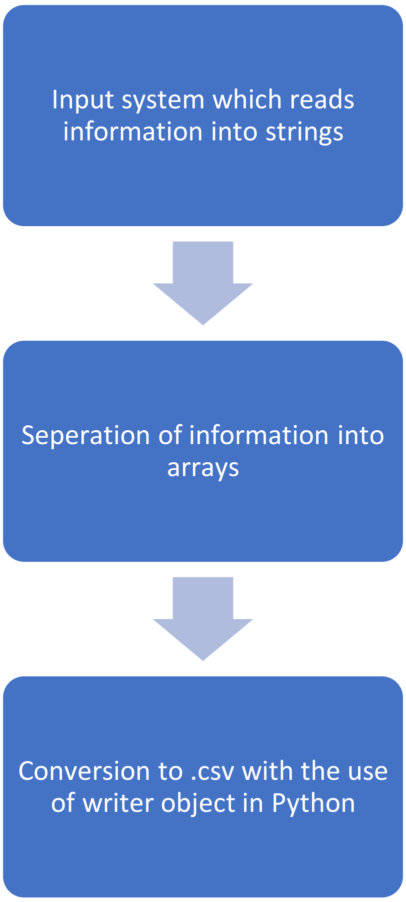


3.1 GEDCOM parser subsystem

The format used in GEDCOM files as seen in figure 1 is hard to understand and read. Therefore, to achieve a more reader friendly format, a GEDCOM to CSV parser was made which converted the information in the GEDCOM file into a CSV file. A more detailed result of the outcomes of using the GEDCOM parser will be shown in the results section.

Throughout the project, the project team reviewed various parsers for GEDCOM files that are available such as the GEDxlate parser. The investigations into the parsers that are available showed that the existing options are either too isolated to be used in a combined program, or unsuited of the project needs. Therefore, the project team came to the conclusion of creating a custom GEDCOM parser that fulfils the needs of the project and that is able to be combined into the whole program without any errors. The basic principles on how the parser has been made is discussed below.

The parser has been created firstly by implementing an input system that intakes the GEDCOM file and reads the data in the GEDCOM file and then reads them into strings in Python. Once the following component was functional, the separation of the information which was in the GEDCOM file into arrays was done with the use of checking the types of tag (FAM vs INDI, etc) and with the use of the whitespaces in the GEDCOM file. The writer object in Python can then be used to write the separated values from the GEDCOM files into a .csv file. The following .csv conversion is performed with the use of the writerow function, using from each individual event (birth, death, etc.) for each row. A summary of the whole process of the basic principles which were used to create the parser has been shown in the diagram below.

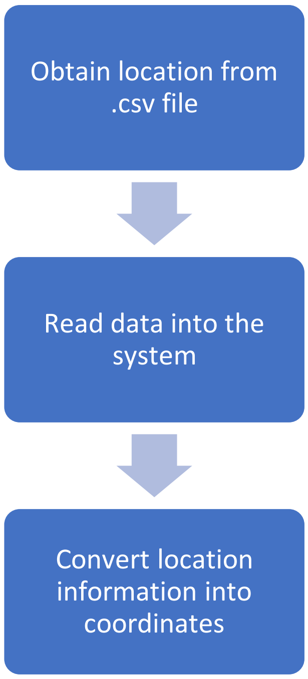


An example of the resulting output from the GEDCOM parser will be shown in appendix B. Moreover, the resulting GEDCOM parser outcomes for various lengths of inputs will also be shown in appendix C and D. Moreover, the specifications document in appendix a was used to understand the format of GEDCOM files and aided in implementing the GEDCOM to .csv parser.

3.2 Geocoding subsystem

The geocoding subsystem is needed to convert the location data into coordinate data. As the GEDCOM files has information such as the location of birth for each individual present in the GEDCOM file the conversion of the locations is required. This was done accordance with the locations listed in the .csv file which were obtained from the GEDCOM files. For the following program, the solution was to use the Google geocoding API. The main programming language used in this subsystem was JavaScript.

The basic main principles in the geocoding API consists of obtaining the relevant location information from the .csv file according to each individual listed. Once this information is obtained, with the use of the API the locations can be converted into coordinates. A summary of the following short process is shown below.



A main challenge which was faced in this process was potential inaccuracy of the Geocoding Subsystem if the API assumed locations when the information provided was not specific enough. An example of this would be, given the input "Melbourne, United Kingdom" the geocoding system would successfully give the coordinates for the city of Melbourne in England. However, in the case where the location information is just given by "Melbourne" the API would instead give the coordinates of Melbourne, Victoria as this is considered the larger city. Two solutions were looked into for this issue. The first solution considered was to implement the Google Place Autofill API alongside the Google Geocoding API. This new API allows for users to interact with the system when ambiguity is detected and select the location from a list of options. While this system allowed for the greatest accuracy as discussed in Section 4.1, and was the most future proof solution, it vastly increased the time the system took to complete the Geocoding as the Place API is somewhat too cautious and had flagged many instances where alternatives technically existed but the default assumption was correct, essentially adding unnecessary time and requiring the user to interact much more with the system. The final issue with this API that caused it to be deemed unusable was the cost. This API was more than three times the price per use versus the Geocoding API and would thus severely increase the operating cost of the system. As such an alternative solution was investigated. This involved checking through the formatting of

GEDCOM files to assess the standardisation of place information. From looking through the GEDCOM 5.5.5 Specifications, given in Appendix A the format of the location information is always given as the depth level in the GEDCOM ("2" for locations), followed by "PLAC" followed by jurisdictions separated by commas. These jurisdictions are levels of location detail such as neighbourhood, postal code, city, county, state, country, etc. Thus an example of a GEDCOM location line would be:

2 PLAC Cove,Cache County,Utah,United States

From this it can be seen that the level of specificity of the location can be gleaned by checking the number of commas as the depth of the information will always be one greater than the number of commas. This process of checking the number of commas was implemented into the Geocoding Subsystem to eliminate locations which could potentially be too non specific. Through use of the Place API for testing, it was decided that three levels of information was enough for locations to not have any issues with being too broad. While there is no guarantee that this will have 100% accuracy as there could exist GEDCOM entries listing, for example, a country, state and city with multiple cities of the same name existing in that state. This would be ambiguous but was rare enough that it was not found in any of the six hundred and three locations tested using the Place API which had already shown itself to be more sensitive than required. As such any location entries with fewer than two commas present was excluded from the Geocoding process and removed from the final .csv file. One additional change was required to the system to remove leading commas as this style of formatting is sometimes used to designate jurisdictions that information is not known about. Using the example above this could be ",,Utah,United States" if the GEDCOM file had four levels of information declared as the standard in the header of the file but only two levels were known for this entry.

This method of designating the desired jurisdictions in the header of the file is considered best practice but unfortunately is not followed regularly by GEDCOM users and as such was not used in our program to learn the exact levels of information provided despite it being a very useful tool. To make up for the lack of use of jurisdiction designations a system was added to the program to do some basic level of detection for the lowest depth present in the location information. As the formatting requires information to be sorted from smallest jurisdiction to largest only the left most entry was considered. This entry was then evaluated rudimentarily and excluded if it included any variation of the word county, as used in GEDCOM formatting, immediately prior to the first comma. This system, combined with the above system for checking number of commas, was effective at detecting the vast majority of instances when a location did not include depth beyond the county level, as desired by the collaborator of this project Colleen Fitzpatrick. In the six hundred and three locations pulled from GEDCOM files by the program during testing, five entries with only county level specificity were determined to have been passed through to the final system. These were all entries containing Country, State and County information but where the county was listed just by its name, ie. "Cache", not in the much more typical format of "Cache County" or "Cache Co.". To minimize the impact of this the location information is provided in the final map, tied to each marker to allow for manual inspection if required. This is a way of introducing human verification into the system to further increase the accuracy of the system without the drawbacks of how it would be implemented through the Place API. This solution greatly reduces the need for user intervention by vastly reducing the number of county level locations that go through the Geocoding process.

Throughout this process, the Google Geocoding API used for obtaining co-ordinate information costs 5 USD per 1000 uses. This is taken from a free monthly credit of $200 supplied by Google. This means that the program is able to call the API 40,000 times per month without cost. To offset the potential cost beyond the initial free calls, a subscription fee is to be added to the program before it can be used by genealogists outside the project group. To minimize the rate at which calls are made, the system includes a section that checks the location to be geocoded and if this location has already been geocoded by the program the coordinate data is simply copied from the prior element, thus saving a call from being wasted. This process is valuable to reduce the calls done and thus save money but is not perfect as the same location can be written in different ways using different jurisdictions. For example "Cove,Cache County, Utah" and "Cove,Utah,United States" are not recognised as the same location until after the Geocoding API is called and the coordinates are found to match. Still, this system can be incredibly efficient if consistent formatting is used throughout a GEDCOM file. For example, in a smaller GEDCOM file tested with only thirty three events with sufficient information, fifteen events occurred in the same location, Huntington, Cabell, West Virginia. Unfortunately, due to the inconsistent nature of GEDCOM entries, these entries were expressed in four different ways with varying combinations of jurisdictions or format but that still reduced the number of API calls from thirty three to twenty two. While this level of success was not always present in the tests, all tests incurred some level of reduction of calls through the implementation of this system.

To summarise, the subsystem takes the location information from the array created by the parser, does two checks to ensure the location is both unambiguous and detailed, checks if the location has not already been found, calls the Google Geocoding API to obtain co-ordinate data for the location and then adds that coordinate data to the output .csv file.

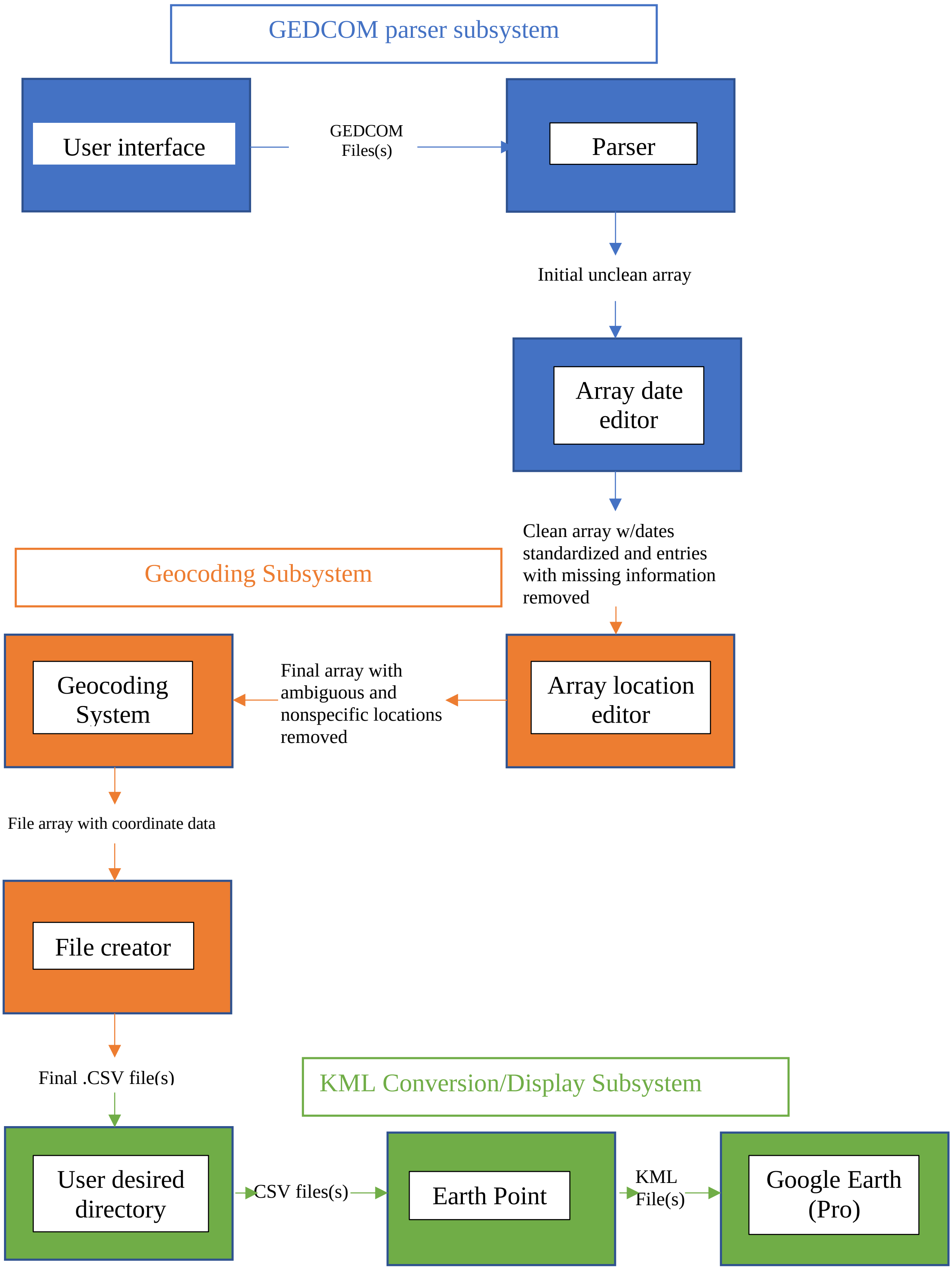
3.3 KML conversion

The conversion to KML was the simplest part of the system to produce as it utilised a commercially available off the shelf solution suggested by the project collaborator Colleen Fitzpatrick. This program, called EarthPoint, had been used by Colleen in the past for this purpose and she highly recommended it saying it covered everything she needed. The project team reviewed the documentation available on EarthPoint's website and confirmed that it had the ability to create the required KML files with all details desired by Colleen including variable icons, worldwide coverage and a time slider to make markers only visible after their event occurred. The resulting KML file is then loaded in Google Earth to visually display the information. Each event has a corresponding marker placed at the coordinates found via the Geocoding Subsystem, the markers are labeled with the individual's name and selecting a marker brings up a description box listing if the event was a birth or death as well as the location name. This can be useful for sanity checking results if something is suspected to be incorrect. If multiple GEDCOM files are converted the markers in each file are given a unique colour, up to twelve files. Beyond this, colours repeat.

3.4 Overall program

With each of the subsystems being produced and tested individually using the output of the prior subsystem, the combining of the systems was quite a quick and painless process. The interactions between subsystems and between their individual component systems are detailed in the diagram below. This diagram also shows the format and description of what was passed between component systems at each step. The processes contained within the first two subsystems are entirely managed by the project program while the final step to convert to KML is done by the user via the EarthPoint program. An example of the final map produced by the system can be seen in figure 3.1. This map contains two hundred and twenty eight markers gathered from four GEDCOM files which each averaged around 1300 lines. The markers are color coded by the GEDCOM file they were extracted from. The four .csv files created from these GEDCOM files can be found in Appendices B through E.

A diagram of the summary of the whole program is shown below.



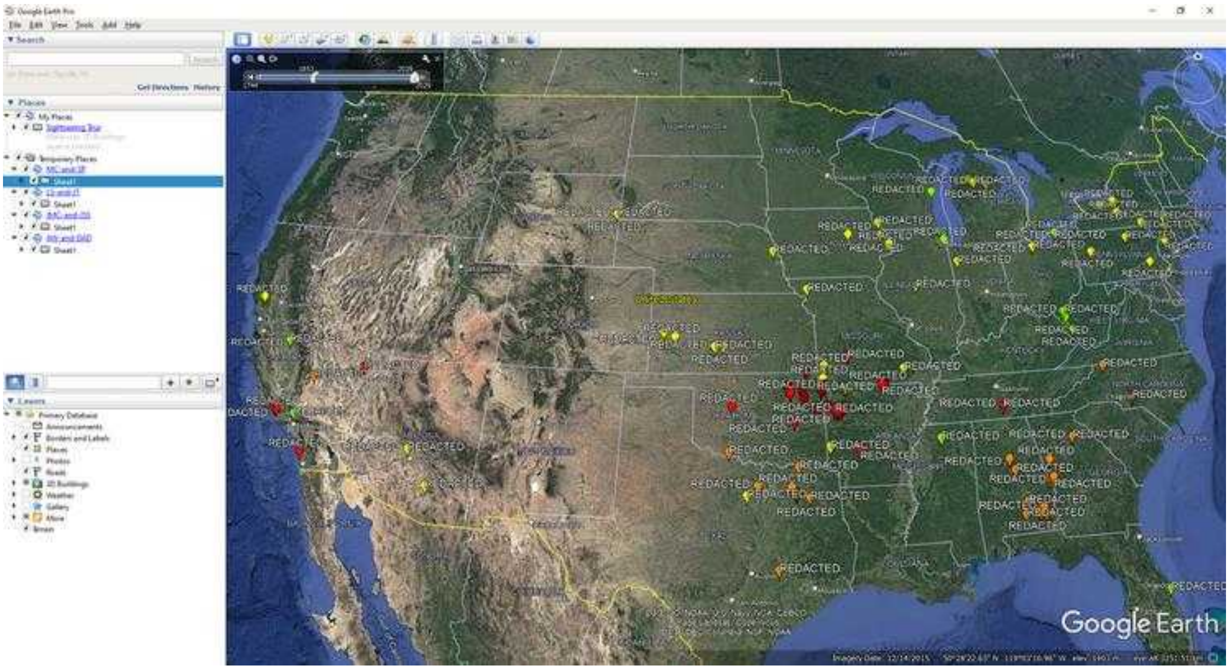


Figure 3.1 Example of final map produced by the whole program

4. Results

As the program subsystems were produced, small tests were conducted to ensure the program was functioning as intended. These tests were done on a small (~1200 lines) GEDCOM file and resulted in a .csv file with thirty three events listed. The GEDCOM file can be found in Appendix F while the resulting .csv can be found in Appendix E.

Note: Due to the files containing personal information about members of the public all names and source links have been redacted from all GEDCOM and .csv files appended in this report.

With the basic tests having been done on each subsystem and all seeming well, the subsystems were combined and more rigorous testing began. This involved running more GEDCOM files through the program to see how the program performed. The program was evaluated using three criteria, accuracy, speed and ease of use. The testing process involved processing over twelve thousand lines of GEDCOM information and resulted in five hundred and ninety events being evaluated as valid information which were then mapped onto Google Earth. The results of these tests are laid out in the following sections.

4.1 Accuracy

When discussions on the requirements of the program were had at the beginning of the project accuracy was repeatedly mentioned as the most important part of the project, as without confidence in the accuracy of the system genealogists would be unable to rely on the information provided and thus the program would be unusable. For this reason the project team devised four metrics to analyse the accuracy of the program and compare that to the prior system used by Colleen Fitzpatrick. These metrics are: Percentage of Ambiguous Locations (%AL), Percentage of Non-Specific Locations (%NL), Percentage of Invalid Entries (%IE) and Percentage of Valid Individuals Missed (%VIM).

%AL describes the percentage of accepted events which through testing were found to have ambiguous location information. This includes entries where the location has multiple options leaving the program to simply pick a single option ("Melbourne" could be in England or Australia, thus ambiguous).

%NL describes the percentage of accepted events which through testing were found to have location information too broad to be 100% accurate. This includes entries where the most specific jurisdiction is a county or larger. These errors were deemed less impactful than %AL errors as the location is still generally correct whereas %AL errors can be off on a global scale.

%IE describes the percentage of accepted events which through testing were found to have either not enough information provided to be useful (Missing a date or name) or where the information provided was not helpful (Placeholder names or dates)

%VIM describes the percentage of valid events present in the GEDCOM file(s) that were found to have been missed by the program or excluded for any non-valid reason.

For all these metrics a lower number is more desirable.

4.1.1 Existing systems

The prior system used by Colleen Fitzpatrick among others involved using at minimum five programs to complete the conversion from GEDCOM to .csv, followed by using EarthPoint. This system was quite good in regards to ensuring no incomplete entries were allowed through as Colleen would be forced to check each entry manually multiple times throughout the process, giving ample opportunity to remove inaccurate or non specific events. Despite these manual checks, partially due to the excessive size of these files, many instances of non specific or broad locations were allowed through, mainly through accepting locations only specific to the county level or, more egregiously, state level. Given that this system had already been in use and Colleen had been happy with its accuracy the project aimed to at minimum match and preferably improve on the accuracy numbers.

This system resulted in the following metrics:

%AL = 0

60 of 417 events listed were of county or state level detail, not city or better.

%NL = 60/417 = 14.4

%IE = 0

417 accepted, 60 inaccurate therefore 357 correctly accepted. 133/490 missed.

%VIM = 27.14

4.1.2 New System

The program functioned as described in Section 3, with a bias toward accepting more data points even if that meant accepting some with county level location detail. This could be adjusted to decrease %NL by increasing the comma requirement in locations from two to three but would cause a higher %VIM. The resulting .csv file used in testing can be found in Appendix G.

%AL = 0

%NL = 69/559 = 12.3

%IE = 0

559 accepted, 69 inaccurate, therefore 490 correctly accepted. 0/490 missed.

%VIM = 0

4.1.3 Comparisons

Table 4.1 shows the accuracy metrics of each system. The data shows that while the new system is not significantly more accurate in regards to only accepting highly detailed locations when compared to the previous system (%NL of 12.3 vs 14.4), it is significantly better at extracting the maximum amount of usable data from the GEDCOM files provided to it. While a significant improvement in accuracy would be desirable the most reasonable solutions discussed all require either more man hours (Manual checks/approvals) or greatly increase the cost of the system (Place API). Free alternatives to the Place API or an increase to the proposed subscription fee could be considered in future work but the program created matches the accuracy of the old system while significantly improving on its ability to extract maximum useable info.

Table 4-1 Accuracy metrics of existing system and new system

Metrics	Existing System	New System
%AL	0	0
%NL	14.4	12.3
%IE	0	0
%VIM	27.1	0

4.2 Speed

When questioning Colleen about her reasons for needing a new system for visualisation the main issue put forth was the extreme effort and time required to convert manually. As such this was the main focus of this project, to create a faster process for genealogists to visualise their data.

4.2.1 Exiting Systems

The current method for visualising GEDCOM files requires many separate programs and manual checks on the .csv file at multiple steps. This is a very time consuming process and is the main reason genealogists such as Colleen do not use visualisation regularly. Colleen has said that depending on the size of the project and number of GEDCOM files this process could currently take her anywhere from an hour to a few days.

4.2.2 New System

The speed of the new system is bottlenecked entirely by the Google Geocoding API. While the API claims to have a limit of fifty requests per second in practice the result has been fluctuating between three and five requests per second. This resulted in the program taking on average two minutes and eighteen seconds to convert a moderately large GEDCOM file (~10,500 lines) into a .csv with all the required information. Three GEDCOM files, each between 1300 and 1500 lines (approximately 4000 lines) were always converted together in under one minute. To estimate the time a GEDCOM file will take to convert, divide the total number of lines in the GEDCOM file(s) by eighty and that will be the approximate conversion time in seconds.

4.2.3 Comparisons

As both systems use EarthPoint for the final KML conversion and GoogleEarth to display the results these timings are the same. Thus the only time difference occurs between the GEDCOM to .csv converting methods. This is the greatest achievement of the new system, drastically reducing the time taken to convert from potentially days to a matter of minutes.

4.3 Ease of Use

The final key issue Colleen had with the existing system was the difficulty she had keeping track of the process as multiple (5+) programs were required for the conversion to .csv alone. Along with the need for manual adjustments the process as a whole needed improvement

4.3.1 Exiting Systems

The existing system is very unintuitive and is also very messy. Therefore, this results in easy to lose track of progress or what is needed to be done next.

4.3.2 New System

The program written in Python was converted to a Windows executable file to lessen the load on the user. Instead of having to install python and all the relevant libraries, prospective users simply need to download a zip file containing the program and its associated files, extract the files and then run the GEDtoCSV windows executable. This executable opens a dialog screen instructing the user to select the GEDCOM file(s) to be converted and allows the user to browse their device for the file via a window very similar to that of file explorer. When the GEDCOM(s) are selected the user is shown a message informing them the process can take time and it will be done when this message disappears. The output files are placed in the same directory as the GEDCOM files are and the user is then free to upload these to EarthPoint as with the old system. This User Interface can be seen in figure 4.1.

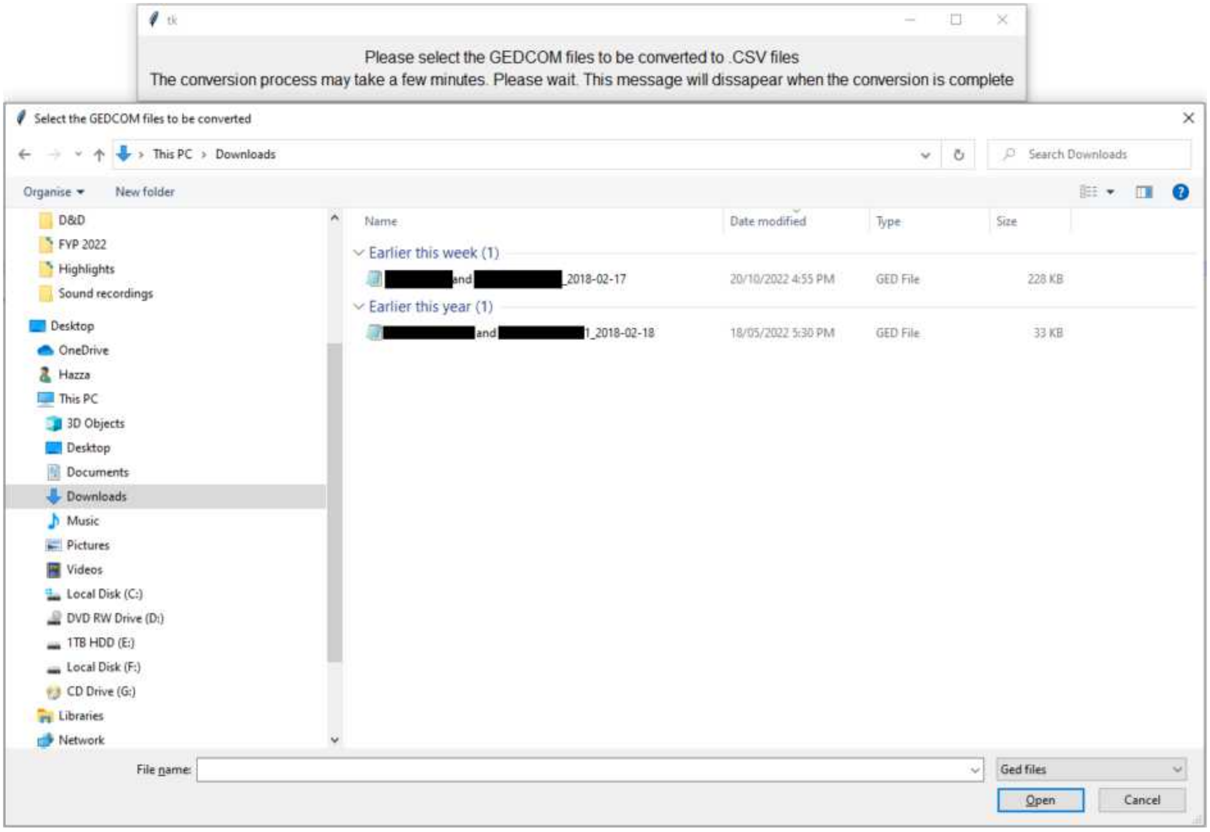


Figure 4.1 User interface as discussed in Section 4.3.2

4.3.2 Comparisons

This process is much smoother in the new program and having tested the new system on non genealogists there have been no issues despite the lack of inherent knowledge on the subject. This User Interface is very simple and leaves no room for error as only GEDCOM files can be selected and every step of the process is explained. A huge improvement over the messy old system.

4.4 Analysis

Overall it can be seen that the tests of the new system show it to be superior in all desired ways. With slight improvements in location specificity and vast improvements in amount of data converted, speed of conversion and ease of use the new system is a great improvement over the old system.

5. Discussion

The following section discusses the results that were obtained from project. Furthermore, how well the project team performed in the project will also be discussed. The section mainly involves the limitations faced in the project, what possible future work can be done with the use of this project and the achievements which the project team has achieved by completing the following project.

5.1 Limitations

The project team observed a few limitations with the completed project. These mainly included the fact that two programs are required to fully obtain the outcome of the project and in terms of the program, it sometimes includes information less specific than a city name, i.e. County name, although still less frequently than the prior methods.

When considering the program requiring two programs to produce all the relevant outcomes, this was due to the time constraint that was put into the project to be finished. To create one whole program for the project instead of using the program EarthPoint for the KML conversion, the project team would have been able to create a Google Earth style map from a .csv file and build upon it to get to a fully functional map. This was deemed unnecessary as the EarthPoint program was sufficient for the needs of the project collaborator Dr Fitzpatrick.

In terms of the information presented in the GEDCOM file, the locations occasionally include information more broad than a city name. The less specific information means that there could be a lot of people populating in that area according to the information in the GEDCOM files which results in it being hard to find a specific person of interest from all the results in the GEDCOM file. The project team were not able to tackle the following limitation as the less specific information was the information that was present in the GEDCOM files and thus could not be enhanced. Perhaps a future system could utilise population density mapping to implement a more accurate marker location in such cases.

5.2 Future work

The project could be extended in the future in a number of various ways. Possibilities of extending the project include:

1. Making a substitute for Earth Point to create a single programs for the entire process
    - As mentioned earlier in the limitations, having two program reduces the time efficiency of the program. Therefore, by creating a single program it can aid in reducing the time taken to obtain the project outcomes.

2. Creating or utilising free geographic lookup tables to eliminate potential operating cost via Google Geocoding API
    - The geographic lookup tables the project team found for the following project have an operation cost therefore, by creating or untilsing free geographic lookup tables it can eliminate potential operating costs associated via Google Geocoding API.

3. Option for user to select from presets for settings if major differences exist in location formatting outside the US (More or less jurisdictions needed for accuracy/unambiguity)

4. More expensive but slightly more accurate version utilising the Google Place API

5. A system to find location information that matches prior entries even if formatted differently to slightly cut down on API calls and therefore cost. Unsure of how this would be implemented, would need more investigation to see if it's feasible.

6. Utilising LineString in EarthPoint to track the movements of individuals (Birth-> Death)
    - This could also tie into using TimeBegin and TimeEnd instead of just TimeBegin. This would create a constantly adjusting map where dots appear and disappear as time is scrolled. This could be a useful addition to the standard option to show all dots as time progresses. Each has their benefits so having both could be a good addition.

The results obtained in the following project, although they are good, there is a lot of improvement that could aid the following project in the future. Therefore, it would be valuable to extend this project in the future, to find various ways to improve the following program and increase its efficiency and accuracy.

5.3 Achievements

When considering the achievements obtained in the following project, it can be said that Objectives 1,2 and 4 were completed at a high standard. Whereas, Objective 3 was completed with the use of the program Earth Point. As mentioned in the limitations section, objective 3 was not made from scratch as expected due to Dr Fitpatrick being familiar with EarthPoint. As such, the collaborator Colleen Fitzpatrick gave the KML conversion least importance for overhaul when comparing the 4 objectives in the project. The research and time required to implement a Google Earth style map from scratch would be high as mentioned in the limitations section. Therefore, the option of using the Earth point program was considered in this case. A more detailed explanation of the achievements obtained for objectives 1,2 and 4 have been discussed in the results section in Section 4 of the report.

6. Conclusion

In summary, the project aim of producing a program capable of taking several GEDCOM files and producing an interactive map displaying the location of the events described in the GEDCOM files was achieved. This was achieved by separating the system into three subsystems, each completing one core objective of the project. Firstly, a system that was capable of taking the information in GEDCOM files and converting that information into .csv files. Secondly, a system that converts location names to coordinate data, making sure to sanity check locations to ensure accurate and reliable results. Finally, a system which generates an interactive map via Google Earth, which places interactive markers at the locations specified by the coordinates in the .csv file was created.In the final system the first two subsystems were created from scratch in Python, utilising the Google Geocoding API for obtaining coordinate data while the third subsystem uitilised EarthPoint to convert the coordinate data into KML files. This method could be adjusted in future to create a custom subsystem for KML conversion inside the single program to save the user some effort but was deemed unnecessary due to the project collaborator, Dr Fitzpatrick, having prior experience with EarthPoint and being more than happy to continue using it. The results of the final program improve on the prior methods of converting from CSV to a visual medium used by Dr Fitzpatrick in three key ways. First, the accuracy, where the amount of events accepted by the system that had location information beyond that of a city level was reduced (11% less) while the total amount of information extracted increased greatly (27% more). Secondly the speed at which conversions could be done was greatly increased, reducing a task that previously took hours or even days down to several minutes. And finally the system is much easier to use relative to the old system due to being more compact and requiring less user intervention and manual work. However, even though the results obtained in the project are good, the system could still be further worked on to develop more targeted methods for use outside the United States and Australia which this project focused on. The system could also likely be condensed down from two programs to one and a more monetarily and time expensive but more accurate system could be developed as an alternative by investigating more robust geocoding solutions and APIs such as the Place API offered by Google.

References

[1] D. Kling, C. Phillips, D. Kennett and A. Tillmar, "Investigative genetic genealogy: Current methods, knowledge and practice", *Science Direct*, p. 1, 2021. Available: http://Investigative genetic genealogy: Current methods, knowledge and practice.

[2]Blau, S, Graham, J, Smythe, L & Rowbotham, S 2020, 'Human identification : a review of methods employed within an Australian coronial death investigation system', *International Journal of Legal Medicine,* vol. 135, no. 1, pp. 375-385.

[3]Hassan, O, Abu, N, Abdin, Z, 2019,' HUMAN IDENTIFICATION SYSTEM:REVIEW', I*nternational Journal of Computing And Business Research,* vol.9 , pp. 1-7.

[4]Ward, J 2019, *How do we identify human remains?*, The Conversation.

[5]Harding, T, Milot, E, Moreau, C, Lefebvre, J, Bournival, J, Vézina, H, … Labuda, D 2020, 'Historical human remains identification through maternal and paternal genetic signatures in a founder population with extensive genealogical record', *American Journal of Physical Anthropology*, vol. 171, no. 4, pp. 645–658.

[6]E. Greytak, C. Moore and S. Armentrout, "Genetic genealogy for cold case and active investigations", *Forensic Science International*, vol. 299, pp. 103-113, 2019. Available: 10.1016/j.forsciint.2019.03.039.

[7]C. Alho, M. Dorn and A. Eduardo, "Would GENEALOMICS be an appropriate term to designate family tree research based on genome-wide data?", *Journal of Genetic Genalogy*, vol. 9, pp. 3-4, 2021. Available: https://jogg.info/wp-content/uploads/2021/12/91-Issue.pdf.

[8] Jones, Tamura. 2010. 'A Gentle Introduction to GEDCOM'.Modern Software Experience.August 24. http://www.tamurajones.net/AGentleIntroductionToGEDCOM.xhtml.

[9] Roued-Cunliffe, Henriette. 2017. "Visualising Historical Networks: Family Trees and Wikipedia". *Academic Quarter | Akademisk Kvarter*, nr. 15 (oktober):40-53. https://doi.org/10.5278/ojs.ak.v0i15.2684.

[10]"Build software better, together", *GitHub*, 2022. [Online]. Available: https://github.com/search?q=gedcom+parser&type=Repositories.

[11] E. Greytak, C. Moore and S. Armentrout, "Genetic genealogy for cold case and active investigations", *Forensic Science International*, vol. 299, pp. 103-113, 2019. Available: 10.1016/j.forsciint.2019.03.039.

[12] Hammer MF, ElisNA (1995) Appendix 1.Y Chromosom ConsortiumNewslett 2:8-9

[13]"What is DNA?: MedlinePlus Genetics", *Medlineplus.gov*, 2021. [Online]. Available: https://medlineplus.gov/genetics/understanding/basics/dna/.

[14]"Allele", *Genome.gov*, 2022. [Online]. Available: https://www.genome.gov/genetics-glossary/Allele.

[15] T. Zerjal et al., "Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis.", *PubMed Central (PMC)*, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1712423/.

[16] S. Luo et al., "Biparental Inheritance of Mitochondrial DNA in Humans", *Proceedings of the National Academy of Sciences*, vol. 115, no. 51, pp. 13039-13044, 2018. Available: 10.1073/pnas.1810946115.

[17]"AncestrySupport", *Support.ancestry.com*, 2022. [Online]. Available: https://support.ancestry.com/s/article/Y-DNA-mtDNA-and-Autosomal-DNA-Tests?language=en_US.

[18] C. Moraes, "Sorting mtDNA Species—the Role of nDNA-mtDNA Co-evolution", *Science Direct*, vol. 30, no. 6, p. 1002, 2019. Available: https://reader.elsevier.com/reader/sd/pii/S1550413119306126?token=2636893951A708F0C0DEA248B5B92DE0A8907F58B0A9D9EAAB73E87D1DD4698E39023715A62B1883ED0511B6F8630084&originRegion=us-east-1&originCreation=20220529074357.

[19] *Ase.tufts.edu*. [Online]. Available: https://ase.tufts.edu/chemistry/hhmi/documents/Protocols/Maternal%20Ancestry_Introduction_Reworked_Aug_25_2011.pdf

[20] C. Kenney, D. Ferrington and N. Udar, *Retina- Chapter 32 - Mitochondrial Genetics of Retinal Disease*. 2013, p. 635.

[21]"Understanding genetic ancestry testing", *UCL Division of Biosciences*, 2022. [Online]. Available: https://www.ucl.ac.uk/biosciences/departments/genetics-evolution-and-environment/research/molecular-and-cultural-evolution-lab/debunking-genetic-astrology/understanding-genetic-ancestry-testing.

[22] J. Pickrell and D. Reich, "Toward a new history and geography of human genes informed by ancient DNA", *Science Direct*, vol. 30, no. 9, p. 381, 2022. Available: https://www.sciencedirect.com/science/article/pii/S0168952514001206.

[23] N. Wyner, M. Barash and D. McNevin, "Forensic Autosomal Short Tandem Repeats and Their Potential Association With Phenotype", *Frontiers in Genetics*, vol. 11, p. 1, 2020. Available: 10.3389/fgene.2020.00884.

[24] *Should we be making use of genetic genealogy to assist in solving crime?*. The Biometrics and Forensics Ethics Group, 2020, p. 10.

[25] "Autosomal DNA testing comparison chart", *International Society of Genetic Genealogy Wiki*, 2022. [Online]. Available: https://isogg.org/wiki/Autosomal_DNA_testing_comparison_chart.

[26] T. Dowdeswell, "Forensic genetic genealogy: A profile of cases solved", *Forensic Science International: Genetics*, vol. 58, p. 102679, 2022. Available: 10.1016/j.fsigen.2022.102679

[27] D. Kennett, "Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes", *Forensic Science International*, vol. 301, pp. 107-117, 2019. Available: 10.1016/j.forsciint.2019.05.016.

[28] T. Callaghan, "Responsible genetic genealogy", *Science*, vol. 366, no. 6462, pp. 155-155, 2019. Available: 10.1126/science.aaz6578.

[29] A. Michael, S. Blatt, M. Isa, A. Redgrave and D. Ubelaker, "Identification of a decedent in a 103-year-old homicide case using forensic anthropology and genetic genealogy", *Forensic Sciences Research*, pp. 1-15, 2022. Available: 10.1080/20961790.2022.2034717.

[30] N. May, "Mystery of Somerton man’s identity solved after 73 years, researchers in Australia say", *The Guardian*, p. 1, 2022.

[31] "Keyhole Markup Language | Google Developers", *Google Developers*. [Online]. Available: https://developers.google.com/kml.

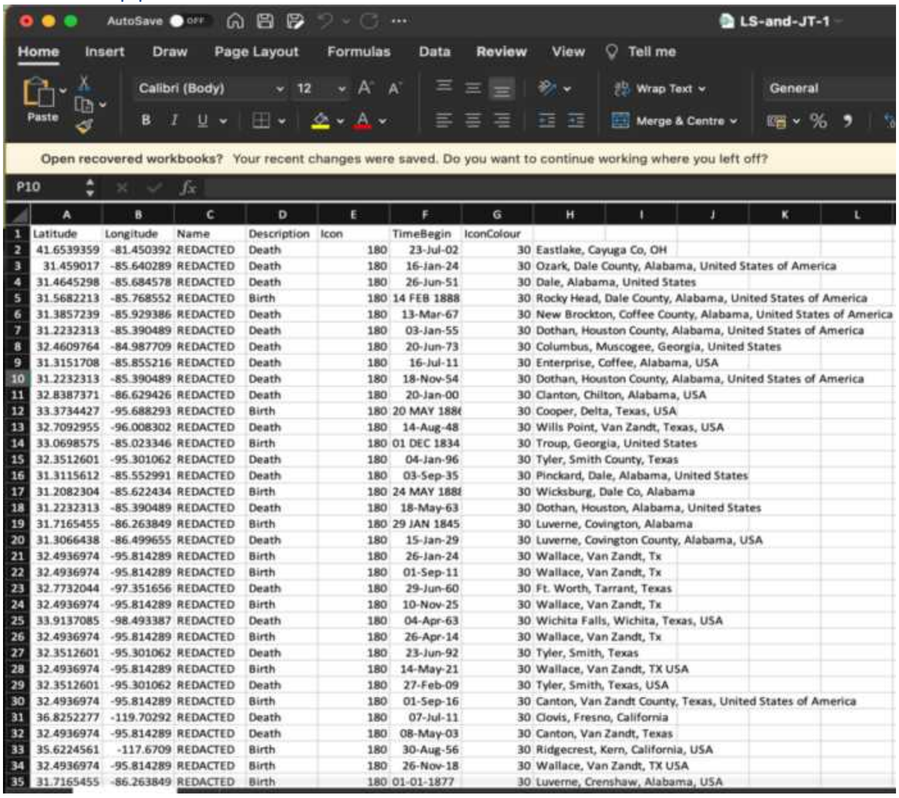
[32] D. De Paor and S. Whitmeyer, "Geological and geophysical modeling on virtual globes using KML, COLLADA, and Javascript", *Computers & Geosciences*, vol. 37, no. 1, pp. 100-110, 2011. Available: 10.1016/j.cageo.2010.05.003.

Appendix A GEDCOM 5.5.5 specs document

The specifications document was used in the project to understand the format of GEDCOM files and in implementing the GEDCOM to .csv parser. This document can be found with the use of the link listed below.

Available at: (https://www.gedcom.org/gedcom.html)

Appendix B- LS and JT .csv



	A	B	C	D	E	F	G	H
1	Latitude	Longitude	Name	Description	Icon	TimeBegin	IconColour	
2	41.6539359	-81.450392	REDACTED	Death	180	23-Jul-02	30	Eastlake, Cayuga Co, OH
3	31.459017	-85.640289	REDACTED	Death	180	16-Jan-24	30	Ozark, Dale County, Alabama, United States of America
4	31.4645298	-85.684578	REDACTED	Death	180	26-Jun-51	30	Dale, Alabama, United States
5	31.5682213	-85.768552	REDACTED	Birth	180	14 FEB 1888	30	Rocky Head, Dale County, Alabama, United States of America
6	31.3857239	-85.929386	REDACTED	Death	180	13-Mar-67	30	New Brockton, Coffee County, Alabama, United States of America
7	31.2232313	-85.390489	REDACTED	Death	180	03-Jan-55	30	Dothan, Houston County, Alabama, United States of America
8	32.4609764	-84.987709	REDACTED	Death	180	20-Jun-73	30	Columbus, Muscogee, Georgia, United States
9	31.3151708	-85.855216	REDACTED	Death	180	16-Jul-11	30	Enterprise, Coffee, Alabama, USA
10	31.2232313	-85.390489	REDACTED	Death	180	18-Nov-54	30	Dothan, Houston County, Alabama, United States of America
11	32.8387371	-86.629426	REDACTED	Death	180	20-Jan-00	30	Clanton, Chilton, Alabama, USA
12	33.3734427	-95.688293	REDACTED	Birth	180	20 MAY 188(	30	Cooper, Delta, Texas, USA
13	32.7092955	-96.008302	REDACTED	Death	180	14-Aug-48	30	Wills Point, Van Zandt, Texas, USA
14	33.0698575	-85.023346	REDACTED	Birth	180	01 DEC 1834	30	Troup, Georgia, United States
15	32.3512601	-95.301062	REDACTED	Death	180	04-Jan-96	30	Tyler, Smith County, Texas
16	31.3115612	-85.552991	REDACTED	Death	180	03-Sep-35	30	Pinckard, Dale, Alabama, United States
17	31.2082304	-85.622434	REDACTED	Birth	180	24 MAY 188	30	Wicksburg, Dale Co, Alabama
18	31.2232313	-85.390489	REDACTED	Death	180	18-May-63	30	Dothan, Houston, Alabama, United States
19	31.7165455	-86.263849	REDACTED	Birth	180	29 JAN 1845	30	Luverne, Covington, Alabama
20	31.3066438	-86.499655	REDACTED	Death	180	15-Jan-29	30	Luverne, Covington County, Alabama, USA
21	32.4936974	-95.814289	REDACTED	Birth	180	26-Jan-24	30	Wallace, Van Zandt, Tx
22	32.4936974	-95.814289	REDACTED	Birth	180	01-Sep-11	30	Wallace, Van Zandt, Tx
23	32.7732044	-97.351656	REDACTED	Death	180	29-Jun-60	30	Ft. Worth, Tarrant, Texas
24	32.4936974	-95.814289	REDACTED	Birth	180	10-Nov-25	30	Wallace, Van Zandt, Tx
25	33.9137085	-98.493387	REDACTED	Death	180	04-Apr-63	30	Wichita Falls, Wichita, Texas, USA
26	32.4936974	-95.814289	REDACTED	Birth	180	26-Apr-14	30	Wallace, Van Zandt, Tx
27	32.3512601	-95.301062	REDACTED	Death	180	23-Jun-92	30	Tyler, Smith, Texas
28	32.4936974	-95.814289	REDACTED	Birth	180	14-May-21	30	Wallace, Van Zandt, TX USA
29	32.3512601	-95.301062	REDACTED	Death	180	27-Feb-09	30	Tyler, Smith, Texas, USA
30	32.4936974	-95.814289	REDACTED	Birth	180	01-Sep-16	30	Canton, Van Zandt County, Texas, United States of America
31	36.8252277	-119.70292	REDACTED	Death	180	07-Jul-11	30	Clovis, Fresno, California
32	32.4936974	-95.814289	REDACTED	Death	180	08-May-03	30	Canton, Van Zandt, Texas
33	35.6224561	-117.6709	REDACTED	Birth	180	30-Aug-56	30	Ridgecrest, Kern, California, USA
34	32.4936974	-95.814289	REDACTED	Birth	180	26-Nov-18	30	Wallace, Van Zandt, TX USA
35	31.7165455	-86.263849	REDACTED	Birth	180	01-01-1877	30	Luverne, Crenshaw, Alabama, USA

	A	B	C	D	E	F	G	H
35	31.7165455	-86.263849	REDACTED	Birth	180	01-01-1877	30	Luverne, Crenshaw, Alabama, USA
36	32.5565244	-95.8633	REDACTED	Death	180	19-Jan-53	30	Canto, Van Zandt, Texas
37	31.3857239	-85.929386	REDACTED	Death	180	01-Nov-86	30	New Brockton, Coffee, Alabama, United States of America
38	36.202858	-81.251883	REDACTED	Birth	180	01-01-1796	30	Wilkes, North Carolina, United States
39	33.7708613	-83.719914	REDACTED	Birth	180	22 MAR 182	30	Walton, Georgia, United States
40	32.4609764	-84.987709	REDACTED	Death	180	28 JUL 1874	30	Columbus, Muscogee, Georgia, USA
41	31.3351716	-85.605213	REDACTED	Birth	180	01 MAR 186	30	Newton, Dale, Alabama, United States
42	31.459017	-85.640289	REDACTED	Death	180	16-Jul-39	30	Ozark, Dale, Alabama, United States
43	32.3552497	-85.189405	REDACTED	Birth	180	01-01-1840	30	Russell, Alabama, United States
44	31.4645298	-85.684578	REDACTED	Death	180	01-01-1879	30	Dale, Alabama, United States
45	31.4645298	-85.684578	REDACTED	Birth	180	09 OCT 1864	30	Dale, Alabama, United States
46	31.3115612	-85.552991	REDACTED	Death	180	10-Oct-36	30	Pinckard, Dale, Alabama, United States
47	31.4645298	-85.684578	REDACTED	Birth	180	01 MAY 185	30	Dale, Alabama, United States
48	31.459017	-85.640289	REDACTED	Death	180	30-Jan-25	30	Ozark, Dale County, Alabama, United States of America
49	33.7708613	-83.719914	REDACTED	Birth	180	01-01-1834	30	Walton, Georgia, United States
50	31.4645298	-85.684578	REDACTED	Death	180	01-Nov-02	30	Dale, Alabama, United States
51	31.4645298	-85.684578	REDACTED	Birth	180	20 DEC 1853	30	Dale, Alabama, United States
52	30.0216106	-96.598306	REDACTED	Death	180	02-May-28	30	Shelby, Texas, United States
53	33.7708613	-83.719914	REDACTED	Birth	180	01-01-1830	30	Walton, Georgia, United States
54	31.4645298	-85.684578	REDACTED	Birth	180	14 FEB 1859	30	Dale, Alabama, United States
55	31.459017	-85.640289	REDACTED	Death	180	22-Aug-35	30	Ozark, Dale, Alabama, United States
56	31.7304339	-86.309685	REDACTED	Death	180	05-Nov-03	30	Rutledge, Crenshaw, Alabama, United States
57	33.2442813	-86.816377	REDACTED	Death	180	05-Oct-11	30	Alabaster, Shelby County, Alabama

Appendix C- MC and JT .csv

MC-and-SP-1

Open recovered workbooks? Your recent changes were saved. Do you want to continue working where you left off?

	A	B	C	D	E	F	G	H
1	Latitude	Longitude	Name	Description	Icon	TimeBegin	IconColour	
2	35.4330846	-93.615129	REDACTED	Birth	180	25-Aug-13	0	Hartman, Johnson, Arkansas, USA
3	35.3076626	-93.882806	REDACTED	Birth	180	25 JUN 1896	0	Ratcliff, Logan, Arkansas, United States
4	35.2971909	-94.04374	REDACTED	Death	180	31-Oct-89	0	Charleston, Franklin, Arkansas, United States
5	35.3054526	-93.953263	REDACTED	Death	180	30 NOV 189	0	Branch, Franklin, Arkansas, United States
6	35.2963579	-93.855351	REDACTED	Birth	180	21 JUN 1875	0	Sixmiles, Arkansas, USA
7	36.1610924	-91.909924	REDACTED	Birth	180	01-01-1861	0	Izard, Arkansas, USA
8	35.3054526	-93.953263	REDACTED	Death	180	19-May-54	0	Branch, Franklin, Arkansas, USA
9	36.1610924	-91.909924	REDACTED	Birth	180	14 FEB 1865	0	Izard, Arkansas, USA
10	36.2009127	-94.387434	REDACTED	Birth	180	01-01-1872	0	Logan, Arkansas, USA
11	36.2009127	-94.387434	REDACTED	Death	180	31 JAN 1899	0	Logan, Arkansas, USA
12	36.1610924	-91.909924	REDACTED	Birth	180	01-01-1863	0	Izard, Arkansas, USA
13	36.1703168	-91.775339	REDACTED	Death	180	05-Sep-05	0	Franklin, Arkansas, USA
14	35.1239397	-86.620794	REDACTED	Birth	180	29 APR 1827	0	Hardin, Lincoln County, Tennessee, USA
15	35.3097149	-94.787283	REDACTED	Death	180	20 SEP 1889	0	Cowlington, Le Flore, Oklahoma, USA
16	35.217314	-93.94048	REDACTED	Birth	180	18 AUG 186	0	Chisomville, Logan County, Arkansas, USA
17	35.3097149	-94.787283	REDACTED	Death	180	22-Jan-12	0	Cowlington, Le Flore, Oklahoma, USA
18	34.9054732	-80.187507	REDACTED	Birth	180	01-01-1785	0	Anson, North Carolina, United States
19	36.1703168	-91.775339	REDACTED	Birth	180	01 JUL 1854	0	Franklin, Arkansas, United States
20	35.3054526	-93.953263	REDACTED	Death	180	18-Feb-34	0	Branch, Franklin, Arkansas, United States
21	36.1610924	-91.909924	REDACTED	Birth	180	05 JUL 1852	0	Izard, Arkansas, USA
22	36.6386937	-95.154134	REDACTED	Death	180	01-Oct-22	0	Vinita, Craig, Oklahoma
23	36.1610924	-91.909924	REDACTED	Birth	180	13 JUN 1849	0	Izard, Arkansas, USA
24	35.3054526	-93.953263	REDACTED	Death	180	15 DEC 1887	0	Branch, Franklin, Arkansas, USA
25	36.1610924	-91.909924	REDACTED	Birth	180	21 MAR 185	0	Izard, Arkansas, United States
26	35.3054526	-93.953263	REDACTED	Death	180	1/01/1934	0	Branch, Franklin, Arkansas, United States
27	35.1239397	-86.620794	REDACTED	Birth	180	03 MAY 183	0	Lincoln, Tennessee, USA
28	35.3054526	-93.953263	REDACTED	Death	180	30 MAY 189	0	Branch, Logan, Arkansas
29	35.4714415	-98.355616	REDACTED	Birth	180	10-Mar-10	0	Hinton, Caddo, Oklahoma, United States
30	36.1539816	-95.992775	REDACTED	Death	180	16-Jul-03	0	Tulsa, Oklahoma, USA
31	35.9420431	-95.883324	REDACTED	Death	180	19-Apr-64	0	Bixby, Tulsa, Oklahoma, USA
32	36.1539816	-95.992775	REDACTED	Death	180	18-Jun-09	0	Tulsa, Tulsa, Oklahoma, USA
33	35.9420431	-95.883324	REDACTED	Birth	180	15-Jul-36	0	Bixby, Tulsa, Oklahoma, USA
34	35.9925829	-94.567999	REDACTED	Death	180	02-Apr-94	0	Westville, Adair, Oklahoma, United States
35	35.9925829	-94.567999	REDACTED	Death	180	25-Aug-85	0	Westville, Adair, Oklahoma, United States
35	35.9925829	-94.567999	REDACTED	Death	180	25-Aug-85	0	Westville, Adair, Oklahoma, United States
36	34.9334298	-95.769713	REDACTED	Birth	180	09-Oct-15	0	McAlester, Pittsburg County, Oklahoma, United States of America
37	34.9334298	-95.769713	REDACTED	Death	180	07-May-02	0	McAlester, Pittsburg County, Oklahoma, United States of America
38	36.1539816	-95.992775	REDACTED	Death	180	15-Dec-86	0	Tulsa, Osage, Oklahoma, USA
39	37.2089572	-93.292299	REDACTED	Birth	180	08-Mar-71	0	Springfield, Greene, Missouri, USA
40	36.1539816	-95.992775	REDACTED	Birth	180	1/01/1939	0	Tulsa, Oklahoma, USA
41	35.7247448	-95.342738	REDACTED	Death	180	1/01/2011	0	Haskell, Muskogee, Oklahoma, USA
42	35.2971909	-94.04374	REDACTED	Death	180	23-Apr-08	0	Charleston, Franklin, Arkansas
43	35.4589055	-94.356558	REDACTED	Death	180	15-Nov-76	0	Van Buren, Crawford, Arkansas, United States
44	34.2804923	-119.29452	REDACTED	Death	180	15-Jun-07	0	Ventura, Ventura, California, USA
45	35.29212	-93.729921	REDACTED	Death	180	28-Dec-78	0	Paris, Logan County, Arkansas, USA
46	35.9420431	-95.883324	REDACTED	Death	180	01-Apr-88	0	Bixby, Tulsa, Oklahoma, USA
47	36.2009127	-94.387434	REDACTED	Birth	180	31-Mar-01	0	Logan, Arkansas, United States
48	35.3076626	-93.882806	REDACTED	Birth	180	7 July 1873?	0	Ratcliff, Logan, Arkansas, USA
49	36.2009127	-94.387434	REDACTED	Birth	180	01-Jan-03	0	Logan, Arkansas, United States
50	36.2009127	-94.387434	REDACTED	Birth	180	30 OCT 1893	0	Logan, Arkansas, United States
51	35.2971909	-94.04374	REDACTED	Death	180	19-Mar-89	0	Charleston, Franklin, Arkansas, United States
52	35.9420431	-95.883324	REDACTED	Death	180	10-Aug-96	0	Bixby, Tulsa, Oklahoma, United States
53	35.9420431	-95.883324	REDACTED	Death	180	13-Jul-68	0	Bixby, Tulsa, Oklahoma, USA
54	35.9925829	-94.567999	REDACTED	Death	180	12-May-01	0	Westville, Adair, Oklahoma, USA
55	35.3076626	-93.882806	REDACTED	Birth	180	25 OCT 1855	0	Ratcliff, Logan, Arkansas, USA
56	35.3054526	-93.953263	REDACTED	Death	180	01-Apr-04	0	Branch, Franklin County, Arkansas, USA
57	34.1975048	-119.17705	REDACTED	Death	180	26-Aug-06	0	Oxnard, Ventura, California, USA
58	36.2009127	-94.387434	REDACTED	Birth	180	20 DEC 1891	0	Logan, Arkansas, United States
59	36.1539816	-95.992775	REDACTED	Death	180	03-Dec-82	0	Tulsa, Oklahoma, United States
60	32.8328112	-117.27127	REDACTED	Death	180	25-Jan-06	0	La Jolla, San Diego, California, USA
61	36.171563	-115.1391	REDACTED	Death	180	19-May-88	0	Las Vegas, Clark, Nevada, USA
62	36.2009127	-94.387434	REDACTED	Death	180	26-Dec-15	0	Logan, Arkansas, United States
63	36.2009127	-94.387434	REDACTED	Birth	180	02 MAY 188	0	Logan, Logan, Arkansas, United States
64	35.9420431	-95.883324	REDACTED	Death	180	04-Apr-65	0	Bixby, Tulsa, Oklahoma, USA
65	33.9407693	-93.140907	REDACTED	Death	180	01-01-1872	0	Greenville, Clark County, Arkansas, USA
66	35.9595442	-95.369412	REDACTED	Death	180	29-Mar-03	0	Wagoner, Wagoner, Oklahoma, USA

MC-and-SP

Appendix D- JMC and JSS .csv

	A	B	C	D	E	F	G	H
1	Latitude	Longitude	Name	Description	Icon	TimeBegin	IconColour	
2	41.6539359	-81.450392	REDACTED	Death	180	23-Jul-02	60	Eastlake, Cayuga Co, OH
3	32.4488736	-97.790576	REDACTED	Death	180	28-Feb-17	60	Granbury, Hood County, Texas, United States of America
4	40.8693272	-74.663464	REDACTED	Birth	180	01-01-1792	60	Roxbury Twp, Morris Co., New Jersey
5	41.1219995	-83.015747	REDACTED	Death	180	01-01-1873	60	Republic, Seneca Co., OH
6	37.6439071	-98.737591	REDACTED	Death	180	02-Oct-13	60	Pratt, Pratt County, Kansas, USA
7	36.0504981	-93.518242	REDACTED	Death	180	31-Jul-43	60	Kingston, Madison, Arkansas
8	44.1025148	-85.146984	REDACTED	Birth	180	24-Jul-16	60	Marion, Osceola, Michigan
9	38.4015781	-122.36081	REDACTED	Death	180	18-Nov-94	60	Yountville, Napa, California, USA
10	44.1025148	-85.146984	REDACTED	Death	180	03-Sep-39	60	Marion, Osceola, Michigan, USA
11	42.3401767	-77.033418	REDACTED	Birth	180	01-01- 1830	60	Orange, New York, USA
12	41.5034271	-74.010418	REDACTED	Birth	180	05 FEB 1852	60	Newburgh, Orange, New York, USA
13	40.692272	-87.276863	REDACTED	Death	180	19-Mar-39	60	Wadena, Benton, Indiana, USA
14	54.3502798	-6.652792	REDACTED	Birth	180	26 NOV 181	60	Armagh, Armagh, Northern Ireland
15	41.9270367	-73.997361	REDACTED	Death	180	05 JUN 1889	60	Kingston, Ulster, New York, United States
16	41.4771349	-74.128494	REDACTED	Birth	180	29 OCT 1867	60	Little Britain, Orange Co, New York
17	42.8142352	-76.809122	REDACTED	Birth	180	29 APR 1812	60	Fayette, Seneca, New York, United States
18	41.5236437	-90.577637	REDACTED	Death	180	07 MAY 188	60	Davenport, Scott, Iowa, United States
19	42.8308007	-104.12272	REDACTED	Birth	180	17-Apr-00	60	Kirtley, Converse, Wyoming
20	42.8142352	-76.809122	REDACTED	Birth	180	20 MAR 182	60	Fayette, Seneca, New York, USA
21	41.9627392	-92.576753	REDACTED	Death	180	06 JUN 1890	60	Tama, Tama, Iowa, United States
22	41.8957751	-90.962952	REDACTED	Birth	180	30 AUG 186	60	Massillon Twp., Cedar Co., Iowa
23	36.0504981	-93.518242	REDACTED	Death	180	20-Dec-03	60	Kingston, Madison Co., Arkansas
24	42.8142352	-76.809122	REDACTED	Birth	180	29 MAR 181	60	Fayette, Seneca, NY
25	Davenport, Scott, IW		REDACTED	Death	180	14 DEC 1868	60	Davenport, Scott, IW
26	42.8308007	-104.12272	REDACTED	Birth	180	15 SEP 1898	60	Kirtley, Niobrara, Wyoming
27	36.4222948	-94.451884	REDACTED	Death	180	30-Mar-77	60	Gravette, Benton, Arkansas, United States
28	41.5034271	-74.010418	REDACTED	Birth	180	01 SEP 1856	60	Newburgh, Orange, New York, USA
29	44.1025148	-85.146984	REDACTED	Birth	180	17 FEB 1886	60	Marion, Osceola, Michigan, USA
30	44.2519526	-85.401162	REDACTED	Death	180	05-Sep-66	60	Cadillac, Wexford, Michigan, USA
31	42.8142352	-76.809122	REDACTED	Birth	180	29 AUG 182	60	Fayette, Seneca, New York, United States
32	42.8142352	-76.809122	REDACTED	Death	180	12 SEP 1820	60	Fayette, Seneca, New York, United States
33	41.2033216	-77.194525	REDACTED	Birth	180	01-01-1787	60	Hard Scrabble, Northumberland Co., Pennsylvania
34	42.7256184	-77.874169	REDACTED	Death	180	01-01-1835	60	Mount Morris, Livingston Co., New York
35	42.8142352	-76.809122	REDACTED	Birth	180	15 NOV 181	60	Fayette, Seneca, New York, United States
36	41.122	-83.01575	REDACTED	Death	180	24 FEB 1842	60	Republic, Seneca, Ohio, United States
37	41.789412	-75.69062	REDACTED	Birth	180	25 NOV 178	60	Scrabble, Hard county, PA, USA
38	41.122	-83.01575	REDACTED	Death	180	13 MAY 184	60	Republic, Seneca County, Ohio
39	42.814235	-76.80912	REDACTED	Birth	180	31 JUL 1826	60	Fayette, Seneca County, New York
40	42.445002	-90.93208	REDACTED	Death	180	19-Jun-08	60	Epworth Dubuque, Iowa, USA
41	40.712152	-79.72877	REDACTED	Birth	180	22 NOV 174	60	Buffalo, Butler Co., Pennsylvania
42	42.904788	-76.86274	REDACTED	Death	180	16 AUG 183	60	Waterloo, Seneca Co., New York
43	36.050498	-93.51824	REDACTED	Birth	180	15-Mar-04	60	Kingston, Madison, Arkansas
44	34.036783	-94.34146	REDACTED	Death	180	06-Feb-86	60	Dequeen, Sevier, Arkansas
45	41.203322	-77.19452	REDACTED	Birth	180	01-01-1781	60	Hard Scrabble, Northumberland Co., Pennsylvania
46	43.980855	-84.4864	REDACTED	Death	180	24-Feb-46	60	Gladwin, Gladwin, Michigan, USA
47	42.814235	-76.80912	REDACTED	Birth	180	13 MAY 181	60	Fayette, Seneca, New York, United States
48	41.311028	-96.34656	REDACTED	Death	180	12 JAN 1876	60	Valley, Douglas, Nebraska, United States
49	42.830801	-104.1227	REDACTED	Birth	180	31-Jan-02	60	Kirtley, Converse, Wyoming
50	36.050498	-93.51824	REDACTED	Death	180	14-Jan-05	60	Kingston, Madison, Arkansas
51	41.203322	-77.19452	REDACTED	Birth	180	10 APR 179	60	Hard Scrabble, Northumberland Co., Pennsylvania
52	41.122	-83.01575	REDACTED	Death	180	21 JUN 1831	60	Republic, Seneca Co., Ohio
53	41.203322	-77.19452	REDACTED	Birth	180	17 FEB 1783	60	Hard Scrabble, Northumberland Co., Pennsylvania
54	42.867984	-76.98556	REDACTED	Death	180	24 FEB 1832	60	Geneva, Seneca Co., New York
55	41.476761	-74.02375	REDACTED	Death	180	01-01-1890	60	New Windsor, Orange, New York, USA
56	42.340177	-77.03342	REDACTED	Birth	180	01-01-1848	60	Orange, New York, USA
57	41.503427	-74.01042	REDACTED	Death	180	26-Dec-09	60	Newburgh, Orange, New York, USA
58	41.203322	-77.19452	REDACTED	Birth	180	01-01-1789	60	Hard Scrabble, Northumberland Co., Pennsylvania
59	42.904788	-76.86274	REDACTED	Death	180	11 JUN 1855	60	Waterloo, Seneca Co., New York
60	41.122	-83.01575	REDACTED	Death	180	13 AUG 185	60	Republic, Seneca County, Ohio
61	39.767458	-94.84668	REDACTED	Death	180	09-Sep-19	60	St. Joseph, Buchanan County, Missouri
62	36.692845	-90.8294	REDACTED	Birth	180	10 OCT 189	60	Ripley, Missouri, United States
63	32.253979	-110.9742	REDACTED	Death	180	02-Nov-59	60	Tucson, Pima, Arizona, United States
64	38.140216	-101.3885	REDACTED	Birth	180	21-Dec-39	60	West Hibbard, Kearny, Kansas, United States
65	37.619463	-99.1065	REDACTED	Birth	180	1/01/1945	60	Haviland, Kiowa, Kansas
66	36.050498	-93.51824	REDACTED	Birth	180	14-Jul-15	60	Kingston, Madison, Arkansas
67	36.868955	-94.368	REDACTED	Death	180	01-Jan-02	60	Neosho, Newton, Missouri
68	37.805576	-101.3885	REDACTED	Birth	180	29-May-43	60	Kendall, Kearny, Kansas
69	37.97169	-100.8727	REDACTED	Death	180	03-Jan-44	60	Garden City, Finney, Kansas
70	37.619463	-99.1065	REDACTED	Death	180	29-Jun-14	60	Haviland, Kiowa County, Kansas, United States of America
71	40.037876	-76.30551	REDACTED	Birth	180	16 AUG 174	60	Hanover, Lancaster Co., Pennsylvania
72	42.904788	-76.86274	REDACTED	Death	180	26 AUG 182	60	Waterloo, Seneca Co., New York
73	33.448377	-112.074	REDACTED	Birth	180	24-Nov-21	60	Phoenix, Maricopa, AZ
74	33.448377	-112.074	REDACTED	Death	180	21-Mar-98	60	Phoenix, Maricopa, AZ
75	41.503427	-74.01042	REDACTED	Death	180	01-01-1865	60	Newburgh, Orange, New York, USA

Appendix E- AW and DAD .csv

Latitude	Longitude	Name	Description	Icon	TimeBegin	IconColour	
41.6539359	-81.450392	REDACTED	Death	180	23-Jul-02	90	Eastlake, Cayuga Co, OH
38.4192496	-82.445154	REDACTED	Death	180	11-Aug-24	90	Huntington, Cabell, West Virginia
41.6253116	-87.717549	REDACTED	Death	180	10-Jun-68	90	Midlothian, Cook, Illinois, USA
38.4192496	-82.445154	REDACTED	Death	180	24-Jul-29	90	Huntington, Cabell, West Virginia, USA
38.4192496	-82.445154	REDACTED	Death	180	27-Aug-52	90	Huntington, Cabell County, West Virginia, USA
37.6742682	-82.277363	REDACTED	Birth	180	06-May-24	90	Williamson, Mingo, West Virginia
43.73031	-87.728968	REDACTED	Death	180	04-May-15	90	Plymouth, Sheboygan, Wisconsin, USA
37.8664833	-82.538763	REDACTED	Birth	180	24 MAR 189	90	Inez, Martin, Kentucky, USA
37.6742682	-82.277363	REDACTED	Death	180	06-Jan-60	90	Williamson, Mingo, West Virginia, USA
39.7817213	-89.650148	REDACTED	Birth	180	01-01-1873	90	Springfield, Sangamon, IL
38.3231877	-82.220771	REDACTED	Birth	180	30 AUG 186	90	Salt Rock, Cabell, West Virginia (Birth Record)
38.4192496	-82.445154	REDACTED	Death	180	16-Dec-50	90	Huntington, Cabell, West Virginia (Death Cert.)
38.3989711	-82.578214	REDACTED	Death	180	20-Dec-76	90	Kenova, Wayne County, West Virginia, United States of America
32.4936974	-95.814289	REDACTED	Birth	180	01-Sep-16	90	Canton, Van Zandt County, Texas, United States of America
36.8252277	-119.70292	REDACTED	Death	180	07-Jul-11	90	Clovis, Fresno, California
38.2214748	-82.442375	REDACTED	Birth	180	02 MAR 185	90	Wayne, Wayne, West Virginia, USA
38.4192496	-82.445154	REDACTED	Death	180	14-Feb-05	90	Huntington, Cabell, West Virginia
38.4192496	-82.445154	REDACTED	Birth	180	17-Mar-05	90	Huntington, Cabell, West Virginia (Birth Record)
38.4192496	-82.445154	REDACTED	Death	180	13-Aug-05	90	Huntington, Cabell, West Virginia (Death Record)
38.4192496	-82.445154	REDACTED	Death	180	31-Dec-69	90	Huntington, Cabell, West Virginia, USA
38.1598501	-82.475276	REDACTED	Birth	180	17 JUN 1882	90	Stonewall, Wayne County, West Virginia, USA
38.4192496	-82.445154	REDACTED	Death	180	15-Aug-67	90	Huntington, Cabell County, West Virginia, United States of America
38.4192496	-82.445154	REDACTED	Death	180	22-Feb-67	90	Huntington, Cabell, West Virginia (Death Record)
34.1514989	-89.631474	REDACTED	Birth	180	06 JAN 1888	90	Water Valley, Mississippi, USA
38.4192496	-82.445154	REDACTED	Death	180	25-Mar-50	90	Huntington, Cabell, West Virginia (Death Cert.)
34.1477849	-118.14452	REDACTED	Death	180	16-Dec-51	90	Pasadena, Los Angeles, California, USA
28.0716832	-80.653388	REDACTED	Death	180	12-Jun-98	90	West Melbourne, Brevard, Florida, USA
38.4192496	-82.445154	REDACTED	Death	180	10-Jan-64	90	Huntington, Cabell County, West Virginia, United States of America
38.4192496	-82.445154	REDACTED	Death	180	28-May-01	90	Huntington, Cabell, West Virginia USA (Cemetery Records)
38.4048042	-82.600437	REDACTED	Birth	180	27-Jun-04	90	Cattlesburg, Boyd County, Kentucky, USA
41.8781136	-87.629798	REDACTED	Death	180	18-Jan-49	90	Chicago, Cook Co., IL
38.4192496	-82.445154	REDACTED	Birth	180	24-Feb-12	90	Huntington, Cabell, West Virginia
38.4192496	-82.445154	REDACTED	Death	180	24-Dec-20	90	Huntington, Cabell, West Virginia USA (Cemetery Records)

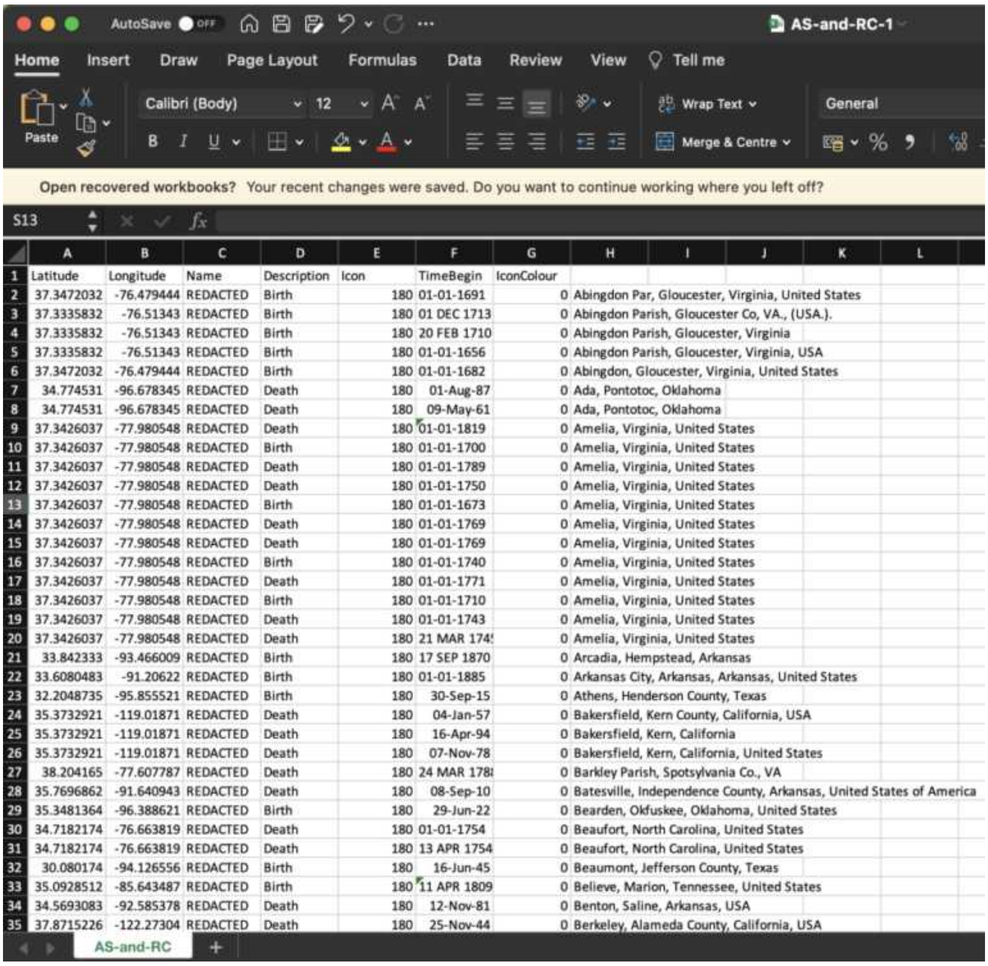
AW-and-DAD

Appendix F- AW and DAD .ged

```
0 HEAD
1 SOUR FTM
2 VERS Family Tree Maker (21.0.0.388)
2 NAME Family Tree Maker for Windows
2 CORP Ancestry.com
3 ADDR 360 W 4800 N
4 CONT Provo, UT 84604
3 PHON (801) 705-7000
1 DEST GED55
1 DATE 18 FEB 2018
1 CHAR ANSI
1 FILE REDACTED
1 SUBM @SUBM@
1 GEDC
2 VERS 5.5
2 FORM LINEAGE-LINKED
0 @SUBM@ SUBM
0 @I1@ INDI
1 NAME REDACTED
1 SEX U
1 BIRT
2 DATE ABT 1935
2 PLAC estim based on age at death
1 DEAT
2 DATE 23 JUL 2002
2 PLAC Eastlake, Cayuga Co, OH
1 OBJE @M1@
1 FAMC @F2515@
0 @I2@ INDI
1 NAME REDACTED
1 SEX F
1 FAMS @F2515@
0 @I5242@ INDI
1 NAME REDACTED
1 SEX F
1 BIRT
2 DATE ABT 1970
2 PLAC per fa birth
1 OBJE @M1339@
1 FAMS @F1304@
0 @I309@ INDI
1 NAME REDACTED
1 SEX M
1 FAMS @F2513@
1 FAMC @F1304@
0 @I4322@ INDI
1 NAME REDACTED
```

The following shows how information is presented in a GEDCOM file at a larger scale. It is to be noted that there are more than 1000 lines of information. Therefore, one screenshot of how the information is presented in a GEDCOM file is shown.

Appendix G- AS and RC .csv



	A	B	C	D	E	F	G	H
1	Latitude	Longitude	Name	Description	Icon	TimeBegin	IconColour	
2	37.3472032	-76.479444	REDACTED	Birth	180	01-01-1691	0	Abingdon Par, Gloucester, Virginia, United States
3	37.3335832	-76.51343	REDACTED	Birth	180	01 DEC 1713	0	Abingdon Parish, Gloucester Co, VA., (USA.).
4	37.3335832	-76.51343	REDACTED	Birth	180	20 FEB 1710	0	Abingdon Parish, Gloucester, Virginia
5	37.3335832	-76.51343	REDACTED	Birth	180	01-01-1656	0	Abingdon Parish, Gloucester, Virginia, USA
6	37.3472032	-76.479444	REDACTED	Birth	180	01-01-1682	0	Abingdon, Gloucester, Virginia, United States
7	34.774531	-96.678345	REDACTED	Death	180	01-Aug-87	0	Ada, Pontotoc, Oklahoma
8	34.774531	-96.678345	REDACTED	Death	180	09-May-61	0	Ada, Pontotoc, Oklahoma
9	37.3426037	-77.980548	REDACTED	Death	180	01-01-1819	0	Amelia, Virginia, United States
10	37.3426037	-77.980548	REDACTED	Birth	180	01-01-1700	0	Amelia, Virginia, United States
11	37.3426037	-77.980548	REDACTED	Death	180	01-01-1789	0	Amelia, Virginia, United States
12	37.3426037	-77.980548	REDACTED	Death	180	01-01-1750	0	Amelia, Virginia, United States
13	37.3426037	-77.980548	REDACTED	Birth	180	01-01-1673	0	Amelia, Virginia, United States
14	37.3426037	-77.980548	REDACTED	Death	180	01-01-1769	0	Amelia, Virginia, United States
15	37.3426037	-77.980548	REDACTED	Death	180	01-01-1769	0	Amelia, Virginia, United States
16	37.3426037	-77.980548	REDACTED	Birth	180	01-01-1740	0	Amelia, Virginia, United States
17	37.3426037	-77.980548	REDACTED	Death	180	01-01-1771	0	Amelia, Virginia, United States
18	37.3426037	-77.980548	REDACTED	Birth	180	01-01-1710	0	Amelia, Virginia, United States
19	37.3426037	-77.980548	REDACTED	Death	180	01-01-1743	0	Amelia, Virginia, United States
20	37.3426037	-77.980548	REDACTED	Death	180	21 MAR 174!	0	Amelia, Virginia, United States
21	33.842333	-93.466009	REDACTED	Birth	180	17 SEP 1870	0	Arcadia, Hempstead, Arkansas
22	33.6080483	-91.20622	REDACTED	Birth	180	01-01-1885	0	Arkansas City, Arkansas, Arkansas, United States
23	32.2048735	-95.855521	REDACTED	Birth	180	30-Sep-15	0	Athens, Henderson County, Texas
24	35.3732921	-119.01871	REDACTED	Death	180	04-Jan-57	0	Bakersfield, Kern County, California, USA
25	35.3732921	-119.01871	REDACTED	Death	180	16-Apr-94	0	Bakersfield, Kern, California
26	35.3732921	-119.01871	REDACTED	Death	180	07-Nov-78	0	Bakersfield, Kern, California, United States
27	38.204165	-77.607787	REDACTED	Death	180	24 MAR 178	0	Barkley Parish, Spotsylvania Co., VA
28	35.7696862	-91.640943	REDACTED	Death	180	08-Sep-10	0	Batesville, Independence County, Arkansas, United States of America
29	35.3481364	-96.388621	REDACTED	Birth	180	29-Jun-22	0	Bearden, Okfuskee, Oklahoma, United States
30	34.7182174	-76.663819	REDACTED	Death	180	01-01-1754	0	Beaufort, North Carolina, United States
31	34.7182174	-76.663819	REDACTED	Death	180	13 APR 1754	0	Beaufort, North Carolina, United States
32	30.080174	-94.126556	REDACTED	Birth	180	16-Jun-45	0	Beaumont, Jefferson County, Texas
33	35.0928512	-85.643487	REDACTED	Birth	180	11 APR 1809	0	Believe, Marion, Tennessee, United States
34	34.5693083	-92.585378	REDACTED	Death	180	12-Nov-81	0	Benton, Saline, Arkansas, USA
35	37.8715226	-122.27304	REDACTED	Death	180	25-Nov-44	0	Berkeley, Alameda County, California, USA

It is to be noted that there are more than 500 lines of information present in the following .csv file. Therefore, only one screenshot has been attached.