Geospatial Family History Research: Designing a Research Tool

A senior project by:

Michael Moore, Chris Hallstrom, Chris Hess, Craig Fryer,

Derek Caswell, Michael Benner, Jared Butterfield, and Richard Helps

Brigham Young University Information Technology

Abstract

One of the core questions in family history and genealogy is “How do I find out more information regarding my ancestors and where they lived?”. Within this question lie other questions and problems such as the time and effort needed to adequately search.

With the advent of easily accessible online geocoding and mapping services available resources can now be searched by their geographical relation to a given point of interest, such as the birth place of an individual or a city name. The Family Tree Mapper project combines information from many available online resources with a map interface to allow users to quickly assess which resources are available and facilitate quicker research.

Developing this tool required a design process incorporating user interface design, geocoding software design, geospatial error analysis and correction and several other aspects. These design considerations are presented in this report, which will proceed by explaining the outline of our project, objectives and methods with an emphasis in user design and testing and geocoding.

1. Introduction

Many aspects of family history are geographically related. This report describes the development of a research tool using geospatial searching linked to genealogical databases. The tool, called “FamilyTreeMapper” (FTM), connects ancestors with information stored in genealogical databases and other resources such as cemeteries, file repositories and also community information, located geographically close to where individuals lived and died. These resources are all presented to the user graphically and can be used for further research and exploration. For example by selecting an individual in a genealogy tree their birth and death places are displayed on a map. In addition family history research centers, cemeteries, tourist information and photos of the area are also available. Developing this tool required a design process incorporating user interface design, geocoding software design, geospatial error analysis and correction and several other aspects.

2. Objectives

FTM provides an important new component in enhancing genealogical research and promoting family history. The users will have an interactive experience with powerful tools to accomplish tasks related to genealogy and family history.

The primary objectives of the project are:

- Help users relate their family history to places and other locale-based information and allow them to visualize their family history by mapping their ancestor's life events.
- Provide access to geographically related resources to help users discover their personal legacies.
- Provide a single access point for many of the resources pertaining to family history.
- Design the system to meet the needs of typical (non-technical) family history researchers.

3. Point and Click Research

FTM is a graphical research tool. Information related to someone in a genealogy tree is obtained by simply clicking on an ancestor's name. The user will load their genealogy into the FamilySearch service (familysearch.org) provided by The Church of Jesus Christ of Latter-day Saints. Once a user's information has been submitted and one or more individuals are selected, some relevant information will be mapped on an interactive map. This will allow users to visually see their ancestors' locations and follow the events in their life such as where they were born, married, or died. A collection of research tools will also be provided which will guide the user's research in a way no current family history product does. FTM centralizes and integrates several tools with a single consistent user interface.

Some of these tools will be locale-based, topic-specific search tools for finding different types of family history information related to a specific place. Many of these are tools developed by third parties that have been integrated into FTM, Such tools are referred to in this report as "plugins".

Family history research has traditionally been tied to place names and thus resources were mostly organized by place name. In recent years due to increased availability of GPS devices and other technological advances, geocoding services have been created to attach any location to its corresponding latitude and longitude. Photos, churches, cemeteries, court houses and many other types of information have been geocoded. With these newly geocoded locations, we can now do geospatial family history research. The power of this approach can best be illustrated in the following example:

> *A family historian knows that their ancestor worked for a mining company in Ironwood, Michigan in the 1850s, but is unsure in which neighboring town they lived. Most towns from that time period have been abandoned and do not exist on current maps. How can this family historian search all surrounding areas?*

With traditional place name research, the family historian would need to find old maps to find the names of the old towns, and then find the locations of churches and cemeteries in each of those towns. With geospatial searching, they could tell the computer “show me all cemeteries and churches within 25 miles of Ironwood, MI”. The computer will then use the longitude and latitude of Ironwood, MI and do a search for all geocoded cemeteries and churches within the requested radius.

4. Genealogy vs. Family History

In order to understand one of the aims of the project, it is important to clarify the difference between genealogy and family history as used in this report. While they are linked to each other and their meanings are intertwined, they have different end results.

Genealogy is the gathering and recording of ancestors and descendants. This study involves the obtaining and recoding names, places and dates of ancestor on Pedigree charts.

Family history on the other hand encompasses genealogy, but also involves studying the stories of one’s descendants. Family history helps to add an empathetic appreciation for the lives of ones ancestors and helps to enrich personal legacies. Often amateur researchers tire of genealogy when no new names or dates surface, but those involved in family history find it more exciting and motivating to continue to seek out information about their ancestors and about themselves.

One of the sub-goals of the FTM project is to help users who are doing only genealogy to bridge over to family history. An additional sub-goal is to help family history researchers be more efficient in their current research.

5. Designing for the User

An essential element of this design is recognizing the (lack of) computer skills of the typical family history user. The system will not be successful unless usability of the target audience is integral to the overall design (Norman 1990, p1, Nielsen 1994, p2-5). For this reason, User Interface design was incorporated into the overall design process.

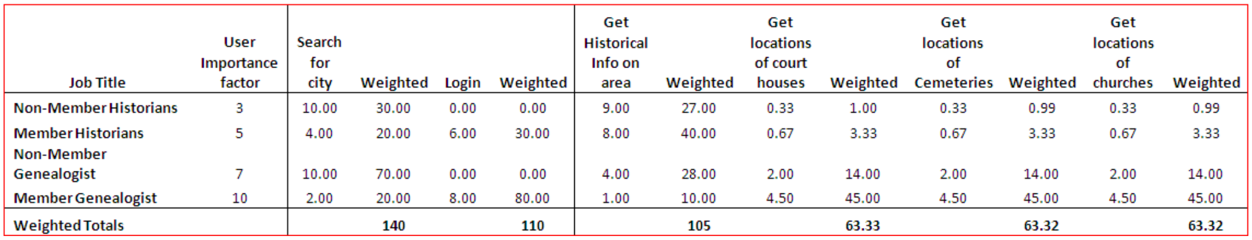
Given the widespread scope of potential users, the selection of appropriate end users of the web service is essential. If the wrong end users are selected, the web service will not meet the goals or objectives of either the users or developers. To locate potential users, demographic data was obtained from the Family History Library in Salt Lake City, a known hub of avid genealogical researchers from around the world. With this data of target users, we evaluated the demographic information and a created a user profile or ‘persona’ for development. Additional interviews were conducted to evaluate the needs and wants of users.

User Benefits

Family Tree Mapper allows users to plot their family history on a simple interactive map, which facilitates in visualizing events of their ancestors and provide location-relevant information. The goal is to interest users in performing more family history rather than just genealogy.

Designing the User Interface

Mockups and low fidelity version of our site were created for user testing. Storyboarding was used to create a simulated experience for users to try various prototypes in a formative evaluation. Additional general information was gathered through observation of the users and observations were analyzed to determine the most frequency performed tasks. The overall layout and design of the web service is based on these tasks. A list of the tasks identified with an indication of their relative frequency is shown below.



Job Title	User Importance factor	Search for city	Weighted	Login	Weighted	Get Historical Info on area	Weighted	Get locations of court houses	Weighted	Get locations of Cemeteries	Weighted	Get locations of churches	Weighted
Non-Member Historians	3	10.00	30.00	0.00	0.00	9.00	27.00	0.33	1.00	0.33	0.99	0.33	0.99
Member Historians	5	4.00	20.00	6.00	30.00	8.00	40.00	0.67	3.33	0.67	3.33	0.67	3.33
Non-Member Genealogist	7	10.00	70.00	0.00	0.00	4.00	28.00	2.00	14.00	2.00	14.00	2.00	14.00
Member Genealogist	10	2.00	20.00	8.00	80.00	1.00	10.00	4.50	45.00	4.50	45.00	4.50	45.00
Weighted Totals			140		110		105		63.33		63.32		63.32

Task Identification

Tasks that are performed most frequently will be the easiest to access and perform. The layout will include these tasks first in priority and less frequent tasks will be further down in priority and less convenient to access in the layout. From the results of these examples and other user testing, the design and function of the web site has been modified to meet the user’s needs and wants.

The design team feels strongly that the process of constantly including the user concerns within the development cycle loops substantially improves final user acceptance of the product and user satisfaction.

Other User Testing Results

Through user testing we found that pop-ups and side menus were more effective then the multiple page format and the user had no real issues in navigating through the site. It was also found that many users wanted to register with the site before they did anything else.

6. Church Resources

FTM is a third party affiliate with the LDS Church on the new familysearch.org project. The affiliate program was organized to “cultivate productive, mutually beneficial relationships with Third-Parties” (Clarke, 2007). Third party affiliates are given access to support and APIs that would allow them to tie into the Church resources such as:

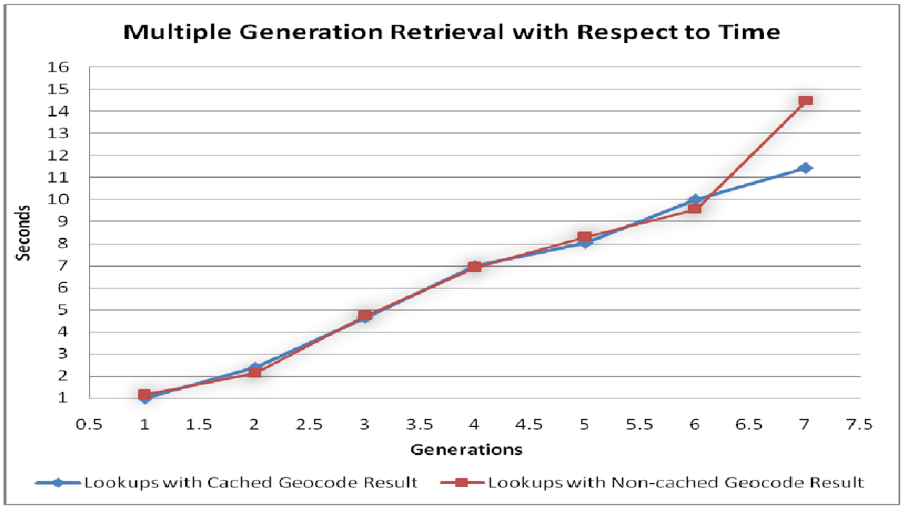
- Geographic Information
    - Provides location Ids
    - Provides geocoding information for most locations
- Feature Requests and Bug Reports

    - Affiliates have the opportunity to interact with the development version of familysearch.org and provide input on the continued development of the project
- Invitations to participate in web conferences/meetings
    - As conferences and meeting occur, invitations will be sent to provide networking and resource building opportunities within the family history field.
- Technical and marketing assistance
    - The church will provide direct support with technical problems as well as to help with marketing and building the usefulness of our project.

With the support of FamilySearch, we can closely tie our product in with another highly used and well known product. There is even an opportunity for continued development into the familysearch.org project.

Working With the Family Search API

The new Family Search provides a web based API which returns data in a standardized (Clarke 2008) XML format. With the API we can fetch information about individuals and the geocodes for many locations. The amount of time it takes to fetch the data the main resource bottleneck in providing the user a good experience. Our goal was to provide the user with a map of four generations, and then allow them to request more generations off of specific branches of their family.



The API is still in Beta, and over the last several months we have seen the performance improve from approximately 12 seconds to fetch four generations, to 7 seconds to fetch four generations. On the left is displayed the results of our latest speed tests, performed on February 28, 2008. In earlier tests, the time to retrieve additional generation grew exponentially. For example it took a half a second to retrieve one generation (2 people), 3 seconds for the two generation (6 people), 15 seconds for the three generation (14 people), 1 minute for four generations (30 people), and over 5 minutes for five generations (64 people).

We have also found that all requests can be made to https://api.familysearch.org with the exception of geocoding with a registered user. Geocoding and only geocoding, requests must be made to http://www.dev.usys.org using a special developer username and password.

7. Plugins

Third party resources or “plugins” for the Family Tree Mapper are a key factor in designing for a successful user experience. They provide tools to enhance genealogical research and promote family history. Here are two specific examples of plugins:

- Family history centers
  - With this plugin, users can easily find family history centers near areas where ancestors lived which provide genealogy and family history related data.

- Cemeteries
  - With this plugin, users can easily find nearby cemeteries to track down important data to further their genealogical research. Cemeteries hold useful information about ancestors including names, dates, and relationships.

Other plugins being developed or evaluated for use in FTM include:

- Flickr - provides images related to the specific area,
- Wikitravel - links to travel information of the area,
- Government Factfinder - provides statistical and census information and
- USGenWeb Project - which is a great resource for genealogical information.

Plugins will behave differently depending on what the plugin type. One type of plugin will plot additional markers onto the map. The other type will provide additional information that is listed in a specific section of the screen assigned to plugins.

Plugin development will first focus on implementing the tools that will have the greatest impact and benefit to the users. There will be a package of plugins for the initial deployment of the project with additional plugins in the works. In addition to developing plugins, an API has been created that allows for any developer to create a plugin that can enhance the usability of the site.

8. Accuracy of Location Information

A major hurdle of FTM is ensuring the accurate retrieval of longitude and latitude information. If a given location is incorrect, it can have serious repercussions. Users will generally accept as fact the program's output, making inaccuracies all the more serious.

Empirical evidence suggests that if a location name is standardized and unambiguous, there is a high probability (98%) of marking the correct place on a map (Whitsel et al, 2006). Unfortunately the same study states that the probability of achieving an unambiguous match ranges from 30% accuracy up to 98% accuracy depending on the vendor. It is possible that results have improved since the study was published in 2006, but we were unable to find more recent data. Further evidence suggests that the mean error distance is less than 2.8 kilometers in rural areas and within 21 meters in densely populated areas (Cayo and Talbot, 2003). Through our own testing, we have found errors and inaccuracies can come from a variety of sources; some of the more common ones are detailed below:

- Same names, different places
  - Ideally each place name would map neatly to a unique geocode, however many locations name's are not unique resulting in a one to many relationship.
  - According to the USGS there are 186 places named 'Riverside' in the United States (USGS, 2008.)In these cases the geocoding service must decide on the best way to resolve the issue.

- No standardized input
  - Convention says that locations should be entered in the format City, County, State, and Country (or equivalently sized political areas) but there is no enforcement of this.

- City Name Changes Over Time
  - e.g. New Amsterdam became New York in the 17th century.
  - Modern map searches do not find this city or similar types of situations.
  - The USGS and other geocoding databases do include some entries which cover these cases; these though are the exception, and not the rule.

- Colloquial / Regional / Abbreviated Names
  - Some locations may be known by unofficial names.
  - e.g. "State-Line" Michigan—this is a known location for recreation, but it is unlikely to appear in a geocoding database because it lacks a substantial permanent population.

- Incomplete coverage
  - No geolocation database has complete worldwide coverage.
  - e.g. When asked why Paraburdoo, Australia (pop. 1600) was not included in the USGS , Marcus W. Allsup of the National Geospatial-Intelligence Agency said:

    "Our policy is to include as many foreign place names as possible. We place special emphasis on countries of high interest or concern to the US, such as those where the military is engaged or relief operations are ongoing or expected. Countries where neither of these conditions is likely to occur are low on our priority list, and Australia would be a very good example of that" (Allsup, 2007).

- Actual Mistakes
  - Any large database relying on human input is going to have a high probability of incorrect entries.
  - Errors in geocoding databases can come from several sources such as misspellings inaccurate numerical input or lack of precision (e.g. accurate to 4 decimal places, or 6).

Partial Solutions

There is no complete solution for location error correction at this time, however there are several automated partial solutions and several solutions requiring user input which could reduce geocoding errors to an acceptable level. Some of these methods follow.

Polling

If three or more databases are available to an application, the application can poll each database to find the supposed geocode for each place name. Any discrepancies could be mitigated through any number of solutions including averaging the three points, flagging the point which differed or more complex accuracy rating systems which would note how often a specific database disagreed with the others and weighing future input from that database by some factor.

Decision Metrics

When incomplete information is presented, the program can use different algorithms can be used to resolve the discrepancies. For example there if just the city “Provo” is listed, the computer will need to decide between Provo which exists in Utah, Serbia, Spain, Bosnia and Herzegovina, Arkansas, and others.

Currently used algorithms include guessing based on the current location of the user (Obtained via GPS or IP address), listing the most populous result, the most searched result, or some combination of the above.

In the case of Family History, we can obtain extra metrics about a location by checking the locations of an ancestor's next of kin. If the ancestor's birth place lists a city and country only, we can check which state and county that individual's parents, siblings and children were born in.

Language Recognition

When the location which is to be geocoded isn't in a standardized format, using natural language interpretation may also increase the probability of finding the correct location. If the location listed is, for example, “Provo or Orem, Utah, Utah, USA”, it is probable that no results will be returned since the city “Provo or Orem” does not exist. By using language recognition, the application could detect that two cities are listed and geocode for one or the other or both, possibly choosing based in part on the metrics listed above.

Automating Discovery of Name Variations

Genealogists have been working for many years to define systems which will allow them to find ancestors when the name has undergone language changes. Some of the more popular methods include the Soundex and NYSIIS algorithms. D. Randall Wilson performed testing on several of these phonetic algorithms and found a recall rate of 93.41% when matching surnames (Wilson, p 4, 2005).

We believe that even higher results are possible with location names since location names can be matched on the city, county, state and country levels in order to help narrow the possible choices (Wilson, 2005).

Correcting the errors

Correcting geocoding errors requires input from users who are familiar with the geocoded location. Getting information from users presents a two faced challenge to database creators. The possibly easier part of the challenge is obtaining the data from the user. The more difficult part of the challenge is determining how reliable the user submitted data is.

User input is potentially very valuable since they are most likely to know the areas where they live. Their input can correct colloquial names, incorrect spellings, missing cities, and most other issues. Accepting user input presents the dilemma of how to verify the submitted information. On one extreme each fact can be verified by a labor-intensive external review process. On the other extreme user submitted data could be assumed to be accurate.

With enough users submitting corrections the users’ own submissions can be used to filter out incorrect user submissions. If five different users say that Provo is located at 40°14′40″N 111°39′39″W and one user says that it is located at 40°45′0″N 111°53′0″W, it is likely that the one person who is different will be the incorrect one. If several submitted corrections are within a certain proximity of each other, the locations could be averaged to offset user error.

9. Security Considerations

Identity theft and related fraud activities become easier when someone's mother's maiden name is revealed. Knowing the names and locations of other living relatives also give a thief clues about locations to look for more pieces of someone's identity. Some potential options for securing the user's data include:

- Not displaying living persons
- Not saving their data, but making them upload it each time
- Saving their data in an encrypted format
- Using SSL with any of the above options

Although the exact route of securing their data has not been finalized, the project has acquired an SSL Certificate and now operates under a secure domain.

10. Upgrade and Support Considerations

As this project is a senior project inside the IT department at BYU, the deadline of the end of Winter Semester 08 is a hard set deadline that will come. As each member of the group is moving on and getting a full time job, this project will be open for others to help as desired.

Plans for Updating - Open-Source Solution

This project code has been licensed as Open Source was a practical decision based on several factors. One consideration was that as an open source project people could contribute plugins and other code to this project. There are many talented developers who may be willing to work with on the project if they have access to the same code base that we do. This project is open for acquisition if someone feels that they could provide value and continued development on it.

11. Summary

FTM has been designed and developed with careful consideration for the needs of the users, the usability of the system, and the technical concerns of the available resources. This project integrates multiple resources to provide a user-friendly, powerful research tool with the goal of helping users perform genealogical research as well as incorporating the more extensive family history resources. It is hoped that this tool will be adopted, used and extended by developers and researchers in the future.

Bibliography

Allsup, Marcus W. "RE: What is the Policy for City Inclusion for Australia." 5 Dec. 2007.

Cayo, Michael R., and Thomas O. Talbot. "Positional Error in Automated Geocoding of Residential Addresses." International Journal of Health Geographics 2 (2003). 28 Feb. 2008 <http://www.ij-healthgeographics.com/content/2/1/10>.

Clarke, Gordon. FamilySearch API Documentation. The Church of Jesus Christ of Latter-Day Saints. Salt Lake City, Utah: FamilySearch Developer Network, 2008. 19 Jan. 2008 <http://devnet.familysearch.org/downloads/documentation/FamilySearchAPI.pdf/view>.

Clarke, Gordon. "Introduction to FamilySearch Affiliate Program." FamilySearch Developer Network. 17 Oct. 2007. The Church of Jesus Christ of Latter-Day Saints. 1 Dec. 2007 <http://devnet.familysearch.org/affiliates/overview/introduction-to-familysearch-affiliate-program/>.

Kirk, Duffin L. "An Area Based Encoding Scheme for Place Names." 1-4. Abstract. Proceedings of 2nd Annual Family History and Technology Workshop (2002).

Leaman, Bob. Genealogical Place Name Normalization. . 3rd Annual Family History Technology Workshop, 3 Apr. 2003, Brigham Young University. 28 Feb. 2008 <http://www.fht.byu.edu/prev_workshops/workshop03/presentations/Bob%20L.ppt>.

Nielsen, J. *Usability Engineering* (Interactive Technologies) (1 ed.). San Francisco: Morgan Kaufmann. (1994).

Norman, D. A. *The Design of Everyday Things* (1st Doubleday/Currency ed.). New York: Doubleday. (1990).

Quass, Dallan. "Identifying Genealogical Content on the Web." 1-3. Abstract. Proceedings of 6th Annual Family History and Technology Workshop (2006).

USGS. "Domestic Names - Frequently Asked Questions (FAQs)." USGS. 03 Dec. 2007. U.S. Department of the Interior. 1 Feb. 2008 <http://geonames.usgs.gov/domestic/faqs.htm>.

Whitsel Et Al. "Accuracy of Commercial Geocoding: Assessment and Implications." Epidemiologic Perspectives & Innovations 3 (2006). 28 Feb. 2008 <http://www.epi-perspectives.com/content/3/1/8>.

Wilson, D. Randall. Name Standardization for Genealogical Record Linkage. 5th Annual Family History Technology Workshop, Apr. 2005, Brigham Young University. 28 Feb. 2008 <http://www.fht.byu.edu/prev_workshops/workshop05/FHTCD/session3/s3-randall_wilson_NameStandardization.pdf>.