Pedigree Generator

Evan L. Ivie
GeneSys and BYU
evan@ivies.org

Peter A. Ivie
GeneSys & University of Illinois
peterivie@gmail.com

1. ABSTRACT

The goal of this project is to automatically connect the residents of Hancock County, Illinois during the years 1850 to 1930 into pedigrees using Federal Census records. This paper describes the steps we have taken to achieve this goal and our results thus far.

2. BACKGROUND

We have made great progress in using computers in genealogical research in the past thirty years. ***Tree Editors*** such as PAF, RootsMagic, Legacy, FamilyTreeMaker and Progeny make it so that we can make changes and additions to our pedigree charts and family group sheets without manually copying whole new sheets. ***Online Depositories*** such as ancestry.com, familysearch.org, archives.com, ellisisland.org have made it so that we can easily search a database for certain names. Tree editors are being enhanced by the variety and richness of the fields that can be entered and the display options. New records and collections of data are being added to online depositories. A recent significant step forward is the ability to connect primary records to individuals in a tree.

Genealogists often follow the following cycle using tree editors and online depositories:

- Study the pedigree in the tree editor.
- Select individual(s) for research.
- Search online depositories for names.
- Discover new information.
- Add the new information to the tree editor.
- Repeat the cycle

As good as this cycle has been in extending pedigrees and in organizing genealogical information, genealogists still struggle with the use of this cycle.

It is difficult to know which parts of a pedigree need work. Queries to online depositories often result in no matches or multiple matches. What should be done then? How might one decide on the level of confidence to have in the data that is found?

One of the reasons for the difficulties we have is that we cannot see the ***context*** of our queries. Context includes temporal and locational proximity, ways to display information, and the frequency of occurrence of names, dates, and places. Mary Williams is an example of the use of context in finding maiden names (see section 6.1 on Maiden Names).

In order to show context we need to advance beyond the current "***search for a name and receive some matches***" approach to genealogical research. Some logic needs to be added so that we can see the context. This paper describes some preliminary ideas on the kind of responses that are needed.

In fact, the vision proposed here is that we should be advancing toward systems that can automatically generate pedigrees – ***pedigree generators***. We would be interested in collaborating with others in setting up a ***National Genealogical Computing Contest*** where pedigree generators compete with each other in

generating the most complete and accurate pedigrees.

2.1 Measures of Relatedness

A key element of our pedigree generator is the use of measures of relatedness. Halbert Dunn, Chief of National Office of Vital Statistics, proposed in 1966 that computers be used to "link" digital records together[1]. Bibliographies on ***record linkage*** can be found in a number of places[2] [3].

In a parallel effort at MIT in 1966 **measures of relatedness** between documents were defined based on information theory[4]. About 26,000 physics research documents were connected together to form clusters of interrelated documents. (As an aside, we note that this effort originated from a desire to be able to connect genealogical records).

2.2 Relatedness in Genealogy

Various teams have investigated using linkage/relatedness techniques to connect genealogical records together over the past 20-30 years. The typical project has taken two adjacent censuses and connected the entries in one census to those in the other census. This has become especially popular since digitized census information is more widely available. Let me cite as an example the work of John Lawson of BYU and David White of Utah State[5] [6].

3. DATABASE

The test database for this study consists of the 1850-1930 censuses of Hancock County, Illinois. We extracted this data from a variety of sources such as familysearch.org, and ancestry.com. Table 1 shows the first family (Wesley Williams family) in the census database.

We also have a database of all of the Ivy/Ivie/Ivies in the US from 1850 to 1930 but did not use it for this experiment.

Census	Given	Mid	Surname	B Year	B Place	City	Gender	Age
1850	Wesley		Williams	1802	Kentucky		Male	48
1850	Ruth		Williams	1804	Vermont		Female	46
1850	Isabell	C	Williams	1833	Illinois		Female	17
1850	Wesley	C	Williams	1834	Illinois		Male	16
1850	Olive	C	Williams	1835	Illinois		Female	15

Table 1. Example of Census Data

4. THE EXPERIMENT

This experiment is designed as a series of steps or passes that process the data. It is designed so that each particular test can start at the beginning running all passes or from an intermediate point if some of the initial passes are not needed. There are currently five passes.

4.1 File Analysis Pass

This pass determines the format of the records and fields in each file. So far we handle "*header files*". A Header file is a file where the first record contains tags describing the fields in subsequent records. We have the code in place to also use Gedcom files but did not use it in this experiment.

4.2 Preprocessing Pass

In this step blank, duplicate, non-Hancock, and meta (header) records are eliminated from the database. Table 2 shows the results.

Census	Fields	Blank	Duplicate	Non-Han.	Meta	Result
1850	13	336	100	464	1	16,570
1860	14	57	583	0	1	27,360
1870	13	0	3,394	0	2	35,950
1880	23	0	213	4,857	2	35,372
1900	25	1,664	0	6,681	6	32,317
1910	20	2	0	0	1	30,714
1920	13	16	9,013	1,211	1	28,550
1930	22	0	724	0	2	25,704
Totals		**2,075**	**14,027**	**13,213**	**16**	**232,537**

Table 2. Record Types in each Census

4.3 Data Cleaning Pass

By standardizing each of the fields in the database we found that we could simplify the

matching algorithms. There is much more we could do in this area, but here are some of the things we have done so far:
Name: Split into given, middle, and surname
Given name: Expand nick names
Given name: Eliminate alternatives
Residence: Split into city, county, and state
Residence: Eliminate "city of", "ward", etc.
Birth year: Eliminate "17 years", "6/12", etc.
Birth state: Convert all abbreviations to full.

4.4 Sorting Pass

Many of the censuses include a family number so individuals can eventually be grouped into families. For those that don't, we sort them into an "as taken" order to help associate families.

4.5 Matching Pass

The approach used initially was to take each record and match it to every other record in the database. A measure of relatedness was then calculated and the record with the top value in each census year that exceeded a threshold was saved as the matching record.

The matching pass actually does two things at once. Individuals are matched with other individuals, and families are matched with other families. Table 5 is an example of the family match output.

Table 3 shows the individual matching results for Charles Goodrich. Note the he covers the entire 80 years and that he was not picked up in the 1870 or 1910 censuses.

Census	Given	Mid	Surname	B Year	Age	B. Place	Township	Relation
1850	Charles	W	Goodrich	1846	4	Illinois		
1860	Charles		Goodrich	1843	17	Illinois	Carthage	
1880	Charles	W.	Goodrich	1843	37	Illinois	Carthage	Self
1900	Charles	W	Goodrich	1843	57	Illinois	Carthage	Self
1920	Charles	W	Goodrich	1844	76	Illinois	Carthage	Self
1930	Charles	W	Goodrich	1844	86	Illinois	Carthage	Head

Table 3. Example of individual matching

Table 4 is an example of Family Matching. Note that the 1870 census was not picked up but was added with a manual search.

Given	Mid	Surname	Age	B Year	B State	Township	Gender	Race	Rel
Wesley	C	Williams	16	1834	Illinois		Male		
1850 CENSUS -- (FAMILY NUMBER: 1)									
Wesley		Williams	48	1802	Kentucky		Male		
Ruth		Williams	46	1804	Vermont		Female		
Isabell	C	Williams	17	1833	Illinois		Female		
Wesley	***C***	***Williams***	***16***	***1834***	***Illinois***		***Male***		
Olive	C	Williams	15	1835	Illinois		Female		
1860 CENSUS -- (FAMILY NUMBER: 1759)									
Wesley		Williams	70	1790	Kentucky	Prairie	Male		
Wesley	***C***	***Williams***	***26***	***1834***	***Illinois***	***Prairie***	***Male***		
Mary	E	Williams	25	1835	Delaware	Prairie	Female		
1870 Census - (Family Number)									
Wesley	*C*	*Williams*	*36*	*1834*	*Illinois*	*Bear Creek*	*Male*	*White*	
Mary	*E*	*Williams*	*34*	*1836*	*Delaware*	*Bear Creek*	*Female*	*White*	
Homer	*G*	*Williams*	*6*	*1864*	*Illinois*	*Bear Creek*	*Male*	*White*	
John	*W*	*Williams*	*5*	*1865*	*Illinois*	*Bear Creek*	*Male*	*White*	
Archibald		*Williams*	*2*	*1868*	*Illinois*	*Bear Creek*	*Male*	*White*	
No Name		*Williams*	*0*	*1870*	*Illinois*	*Bear Creek*	*Male*	*White*	
ohn Crumley)			*17*	*1857*	*Iowa*	*Bear Creek*	*Male*	*White*	
1880 CENSUS -- (FAMILY NUMBER: 4600)									
Wesley	***C.***	***Williams***	***46***	***1834***	***Illinois***	***Prairie***	***Male***	***White***	***Self***
Mary	E.	Williams	44	1836	Delaware	Prairie	Female	White	Wife
Homer	G.	Williams	16	1864	Illinois	Prairie	Male	White	Son
John	W.	Williams	14	1866	Illinois	Prairie	Male	White	Son
Archibald		Williams	12	1868	Illinois	Prairie	Male	White	Son
Robert		Williams	6	1874	Illinois	Prairie	Male	White	Son

Table 4 Example of Family Matching

5 DISCUSSION OF MATCHING

Matching was done at two levels in this experiment: at the individual and at the family level.

5.1 Matching Speed

Using the traditional n-squared approach for the 232,537 records in the database would require n-squared matches: 5.3 billion record matches and about 128 billion field-to-field matches (with an average of 25 fields in each record). Including data cleaning, the creation of auxiliary files, etc. this approach currently takes 3-4 hours. For the number of counties and the size of the US in the 1830-1930 period we estimate that one computer could complete the full US in about one year.
We have further streamlined the process so that

no record is matched with records in a census taken before the birth of that individual. Also in our match we keep the best ten matches for each census, but we are currently using only the best match. We match each record with the other records in its census year to help find duplicates. As an aside we are really excited about a new technique we are exploring that reduces the n-squared algorithm to something closer to an n algorithm.

5.2 What is Best Match?

Exact matching of corresponding fields and summing up the number of matching fields is the simplest way to compare two records. We have used that approach as a benchmark. A slightly more sophisticated approach would be to assign a weight to each field and to sum the weights as was mentioned earlier. A threshold can then be applied to the result to determine if the record should be saved in the best match table[7].

A number of projects have used the EM approach to record linkage. Some have claimed that this eliminates the need for doing a test sample to set the weights. Our goal (unachieved yet) is to use an information theory based approach that uses field value frequencies[8].

6. PARTICULAR CHALLENGES

We include in this paper three examples of particular challenges that the pedigree generator developer faces: maiden names, misspelled surnames, and census accuracy.

6.1 Maiden Names

One of the more challenging components of a pedigree generator is a maiden name algorithm. A woman may have a set of records relating to her before her marriage and a set after her marriage, but with no easy way to connect the two sets and thus connect her pedigree and her descendancy. Marriage records might do this but they do not exist for many areas and many periods of time.

We generally try to find records about a person by doing a name search, but without a surname to match there will generally be too many matches. For example, if we searched for Mary we would get over 500 hits in the Hancock database.

The algorithm we currently favor for census records uses the three fields, the given name, the birth year and the birth state. Consider, for example Mary E. Williams of the previous example. A search for these three fields yields Table 3. Mary Cassingham and Mary Hewitt are 1860 contemporaries and are not candidates for matching with Mary Williams. Thus, the algorithm would suggest that the maiden name might be “Moore” (see 1850 census). This is confirmed in the 1900, 1910, and 1920 censuses (that have additional relationship fields).

Census	Given	MI	Surname	Age	B Year	B State	Residence
1850	Mary	E	Moore	15	1835	Delaware	
1860	Mary		Casingham	27	1833	Delaware	Durham
1860	**Mary**	**E**	**Williams**	**25**	**1835**	**Delaware**	**Prairie**
1860	Mary	E	Hewett	25	1835	Delaware	Wythe
1870	Mary		Cassingham	38	1832	Delaware	Rock Creek
1880	Mary	E.	Williams	44	1836	Delaware	Prairie
1880	Mary		Cassingham	48	1832	Delaware	Prairie
1900	Mary	E	Williams	65	1835	Delaware	Carthage
1910	Mary	E	Williams	75	1835	Delaware	Carthage
1920	Mary	E	Williams	84	1836	Delaware	Carthage

Table 4. Search for “Mary” born in the 1830’s in Delaware

6.2 Misspelled Surnames

Our Pedigree Generator could be much more successful if all of the surnames were accurate. Unfortunately, this is far from the case. There are 59,502 family names in the 8 censuses. There are 11,672 unique family names. Thus, each unique family name occurs an average 5.1 times (18.1 times if you count each occurrence of the name within each family). Of these 11,672 names, there are 5,747 names that only occur

once in the total database.

The set of 5,747 unique names is good place to look for misspelled surnames. We looked in detail at some of them. We found a few cases where the census taker had recorded the names differently than how they are found in other (census) records. We found a number of cases where the handwriting was faded or scribbled making it very difficult to read. We found cases where the extractor did not do a good job of interpreting what was there. And there were a few cases where the family name had evolved (many Hancock County citizens came from Germany, Switzerland, and France).

However, we found several automated techniques for correcting misspelled surnames without manual research. Use of the 3-field Maiden Name Algorithm was very successful in finding and matching up the large variations in surnames. The use of name matching schemes (such as Soundex) was also very useful. Neighbor proximity was also useful but we have not fully implemented it yet.

6.3 Accuracy of Census Data

Census data generally has a bad reputation among genealogists. It is felt that vital, land, court, and other records are much more trustworthy. However, we discovered something unexpected. By pooling the data on a family from several census years we became convinced that the censuses could produce very accurate data. We conjecture that it can be more accurate than most other sources because of the reoccurrence of a given person's information from census to census. And censuses add additional information (temporal evolution of names, residence tracing, etc. that cannot easily be obtained from other sources.

7. RESULTS

The current effectiveness of the pedigree generator is summarized in Table 5. It was able to find a match for 31% of the individual entries. An additional 33% are in families that are matched up through one or more family members. We found no matching entries for the remaining 36% of the entries.

We have watched the matching capability improve as our data cleaning and match algorithm efforts have improved. We guestimate that about 8% of the records are not matched because of births and deaths. For example, anyone under 10 in the 1930 records has no opportunity to be matched with anyone else. Infant mortality was very high during this period of time[9] [10].

We also guestimate that moves into and out of the county might account for 6-8% of the unmatched records. This would include families that lived in the county for only one census 1-19 years). Most of the remaining unmatched records seem to be due to faulty data caused by errors in recording and extraction. One of the tasks on our TO-DO list is to sample the results and come up with better estimates.

Grouping	Number	Percent
Records with 1 or more matches	72,000	31.00%
Records matched through families	77,500	33.30%
Unmatched records	83,200	35.70%

Table 5. Pedigree Generator Match Results

LESSONS LEARNED

We feel that we learned the following things from this experiment:

- Adding family structure to digital databases significantly enhances their utility and ease of use.

- Exposing "context" increases the ability of users to make accurate decisions.
- Utilizing the characteristics of censuses (almost complete coverage, little duplication, temporal reoccurrence) makes decision-making easier and helps display contextual.
- Promising techniques for resolving some of the more difficult problems in genealogical research (maiden name finding, misspelled surnames, faulty data) have been developed.
- By utilizing the data from multiple censuses, the accuracy of census data can be improved.

Finally, we remain convinced that pedigree generators will at some point compete well with professional genealogists.

8. FUTURE WORK

We hope to obtain better estimates of the recall, precision and other results from the Pedigree Generator. We feel that we can improve the matching results by working more on the cleaning and matching algorithms and on the family grouping algorithms.

The layouts of the charts and the structuring of the information are copyrighted. Also, we plan to patent the more innovative processes we have developed. We are anxious that these techniques be widely used. As we refine this system we plan to make licensing available to large-scale consumers in order to fund continued research.

REFERENCES

[1] Halbert L. Dunn. Record Linkage. *American Journal of Public Health,* 36:1412-1414, December 1966.

[2] http://www.recordlink.org

[3] http://www.census.gov/srd/www/reclink/biblio.html

[4] Evan L. Ivie. Measures of Relatedness Between Documents, *PhD Dissertation*, 1-240, MIT Press, 1966.

[5] John S. Lawson. Record Linkage Techniques for Implementing Online Genealogical Research using Census Index Records. *American Statistical Association Section on Survey Research Methods*, 3297-3303, 2006.

[6] John S. Lawson. Find Them in the Census Records. *Genealogical Helper*, 121-129, July/August 2006.

[7] I. P. Fellegi and A. B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64:1183-1210, 1969.

[8] http://www.census.gov/srd/papers/pdf/rrs2004-01.pdf

[9] Evan L. Ivie and Douglas C. Heiner, Deaths in Early Nauvoo, 1839 to 46 And Winter Quarters, The Religious Educator, 163-174, Vol. 10, No. 3 (2009).

[10] Douglas C. Heiner, Evan L. Ivie, and Teresa Lovell Whitehead, Medical Terms Used by Saints in Nauvoo and Winter Quarters, 151-162, Religious Educator, Vol. 10, No. 3 (2009).