Genealogy Research Examples and Alternative Methods

A discussion thread on LinkedIn.com has convinced me that I need to do a much better job on the examples used to explain my concept. Most examples will be real, but a few might be theoretical. The very simple goal of my suggested method is to encourage people to specialize and cooperate on a grand scale. I will provide some examples that suggest people collaborate by geography or by surname, or some combination of the two

In one case, a researcher decided to take all of the records for a small town in Italy, over many generations, and to use the available city records to fit all the names together into family structures. As I recall, there were a total of about 26,000 names covering all generations, which might indicate that the town had perhaps 6000 current residents.

If that was a 10-year task, we might compare it to another 10-year task where one person completed six full generations of her pedigree. For this pedigree operation, that means this lady completed about 128 ancestors.

Comparing the two methods:
26,000 names in Italian city / 128 names from strict pedigree researcher = 203 times more names in same time period.

In other words, the productivity in locating and connecting names was about 200 times as effective in the case of the Italian city as it was for the very dedicated, but otherwise typical genealogy researcher. I am guessing that the researcher in Italy did this part-time and spent about 5000 hours over a 10-year period, meaning that names were recorded at the rate of about five per hour. That is truly a blistering speed, made possible only because all the records for all the generations were probably all sitting in one building.

The main point of this new concept is to get people doing things at which they are extremely efficient, and then when one combines all the separate work products together on a national basis there is another very large boost in overall productivity.

The first example I saw in my studies was the Huff book which contains the descendants of Engelbert Huff, born in 1637 in Norway. The book contains about 15,000 names, of which about 5000 names were people born with the surname of Huff (to a large extent in New York State) who were descendants of this original ancestor. I believe it took about 10 years to put that book together, so if we compare the 5000 family names collected there with the 128 names collected by the lady I mentioned, that means that (5000/128=39) the Huff book, while less efficient than the Italian city project, still had a productivity rate 39 times better than the standard method.

So if we can have researchers who, at the first level, are at least 39 times more efficient than everyone else working today, and maybe up to 200 times more efficient than other workers today, and if we can change the scenario so that EVERYONE is collecting and documenting data from 39 to 200 times more efficiently, then, by extension, the entire process can go 39 to 200 times faster.

Now perhaps we can start to understand the difference between our current "every man for himself" method of doing research, where people only run into cooperative situations by chance, in contrast to a more structured method which involves formal and coordinated cooperation.

If we stick with the geography idea for a minute, there are about 3300 counties in the United States which altogether contain about 330 million people. That would make the average county have 100,000 people. If we assume that there are at least 4 million serious genealogy researchers in the United States, that means that there are, on average, (4 million/3300 =) 1212 researchers in each county. Or we could say that there are about 100 records of deceased people in the country for each of the genealogy researchers (100,000/1,000= 100). We can assume that there are about 330 million deceased people for our country, matching approximately the number of living people in our country. (As a related number, I did some calculations to demonstrate that there were 70 million people who lived and died in the United States before 1930.)

Of course, population and genealogists are not evenly distributed throughout our country, but it would be easy to adjust for these changes in parameters. Hopefully, if there are more people in a county, there will also be more genealogists there.

So, one way to do the entire United States from scratch would be to ask all the county organizations, many of which are already set up, to simply assemble all the historical families in their county using all of the historical records which are available there. So if each of those county genealogists did 100 names, we would have the entire country finished. And it should be relatively easy to do this kind of research, because one could use a process of elimination to put together all the easy families, and then see what is left to fill in the puzzle pieces.

Doing all this in six months time should be a fairly easy. (160 hours per month of work times six months equals 960 hours of work time per volunteer.) For 100 names, that means that people could spend 9.6 hours on each name, and still finish it all in six months. I would guess that it would be a great deal faster than that in most cases, perhaps 10 times faster, meaning that the whole process could be finished in less than a month of work by each participant.

The final step would be to connect the data from each of the 3300 counties together through family generation connections for same-surname situations, and then through marriages where different surname structures are joined together, and the job is done. All the source records have already been assembled and associated with the names, so making the family connections should be relatively easy.

This general process is more or less the way a computer sorting algorithm would work. First it sorts sub-collections of data into the largest possible units that are convenient for its memory constraints, and then it stores those sub-collection units, and then starts merging those sub-collections together into progressively larger collections until the job is finished, and one continuous sorted string of entries is prepared.

Since, by definition, the names that are used are coming from specific documents in that county, the quality of the finished work should be as high as is possible. All of this data would be entered into a database system which is friendly to and supports this kind of cooperative research. (I have a running version of such a software system.)

As a variation to the purely geographical method, we could ask people to take a single surname, such as the Huff surname, except they would substitute one of their own family surnames, and follow it through the necessary geography and time periods as a subset of the total names recorded during that particular geography and time period.

I have emphasized the single-surname, descendent structure method as a way to specialize and narrow each researcher's task, because I thought people would be happy that every name they added to their research structure would be someone who is either an ancestor or a cousin, so they would feel more closely connected. It would also be practical to organize family organizations of all the living same-surname cousins who might be willing to help put all of this together, as was done in the case of the Huff book. And using that single-surname method that would also obviate almost any other kind of coordination processes that would be needed, as might be needed in a county-focused process. But either the county-focused or the surname-focused method would get the job done with similar levels of efficiency, so it is up to the researchers to decide which they prefer.

Statistically speaking, if one begins with an ancient ancestor and proceeds forward in time researching same-surname descendants, typically those families will stay in one area for a long time, making the geographical range of the search for records much smaller. Of course, if someone immigrated somewhere along the line, then there is that jump to be made, but that may be only a very small portion of the total number of names affected. This usually means the range of search for names is much smaller than in the typical pedigree-sequence research techniques, where on any one of 30 different surname lines, the next jump back might be anywhere on the planet.

So what are some of the standard objections to this kind of organization of research work?

1. "It's no fun to just have it done. We want to do this for the rest of our lives without ever being finished, and we only want three generations done, not the 10 to 15 generations this project would put together for us, using all available records. We only care about a few generations for ourselves, and don't care about all the other people in our country."

Our duplication rates could be as high as 37,000 times if we were to thoroughly do an entire country's population using current methods, without the benefit of any totally random, unpredictable, and serendipitous breaks, indicating the potential for a massive waste of researcher time. Cutting the duplication-of-effort rate down to only 100 times because of those serendipitous connections would be helpful, but that is still a huge waste of effort, and a great negative incentive which keeps many people from even trying. Chance encounters and cleverness could indeed reduce that duplication rate, but, more likely, we will simply never do most of the work at all, because it is so difficult and terribly time-consuming if done individually. But deciding to avoid duplication by also deciding to avoid doing any work at all, is not a very good solution.

The point of my method is to subdivide the task and very consciously and deliberately do EVERYTHING ONLY ONCE, with essentially zero duplication, using all available records, and then connect it all together, rather than do small portions of the total task, in a haphazard way, and do them hundreds of times over, or not at all. I believe the LDS family history database contains some names which appear there 10,000 times because of the past inability to avoid duplication of researcher effort and duplication of posted results. As far as I can tell, this vast coordination problem has still not been solved in any rigorous and predictable way.

Obviously, once all the names were done and interconnected, then any interested family members could quickly check all of the source documents and the interconnections of individuals to verify and reverify 100 times, if they wish, all the work that had been done in this first initial push. Corrections could be suggested and made. When this basic network of names had been completed, then people could go on and spend all their time locating pictures and stories and other local and time-dependent data to add to the names of interest to themselves.

The difficulty of individual research, meaning it takes an entire lifetime of dedicated work to follow a pedigree-sequence method back five full generations, plus the vast likelihood of astronomical levels of mostly redundant duplication-of-work using our current uncoordinated methods, means that it is mathematically impossible to ever actually finish all the genealogical connections for the United States, for example, so we could never even focus on another country or area such as Europe, with the goal of totally finishing it. Only by changing the sequence of research processes is there any hope of finishing this task. Even allowing 1000 years to finish the task is not enough under current conditions. It requires an infinite time frame using current methods, or it requires every individual researcher to live 1000 years to finish their pedigree research task back 10-15 generations where the records end. Comparing our current methods with the methods I am suggesting, indicates that the amount of effort that has been put into genealogy research processes in just the past 15 years could have completed the entire United States from scratch 450 times, and still we have barely scratched the surface of the 10 to 15 generations of data remaining to be harvested and placed into family structures.

Our current lifetime constraint on finishing genealogy

One of the characteristics of our current way of doing research, is that, typically, one person starts on a pedigree-sequence research project and tries to finish it all himself or herself. In most cases, all the time one person has in one lifetime to finish this pedigree-sequence research project is about 10,000 hours, perhaps spread over 10 years which is only enough to complete a full five generations of work. That is only 64 ancestors to be found. I hope that most people would like to have 10 generations of research done, which, in many cases, would get the research back as far as there are existing records. (In some cases the records go back 15 generations, but I will focus on the 10 generation situation.) So that automatically means that essentially no one will ever finish 10 generations back on their own, and can never even think about doing a full 15 generations. As I calculate it, it would take 32 times as much effort to finish 10 full generations as it does to finish five full generations, because there are 32 times as many names (2048) involved to be found, and since no one lives 32 lifetimes these days, that initial effort can never be finished.

Let's say that a researcher who has retired from the regular workforce can spend 1000 hours per year on genealogy research, so that it would take about 10 years to put in the necessary 10,000 hours to finish five full generations of research, resulting in 64 ancestors found. In order to finish the research back a full 10 generations, it would take 32 times as long, meaning that it would take about 320,000 hours to finish the work or about 320 years of work. So, to finish the process using our current methods and assumptions and assignments, all we need to do is have individual researchers live to be about 400 years of age, and not be involved in any regular economic work activity, but can spend at least half time working on genealogy research.

Since our current methods are critically tied to how long a single person can spend on their pedigree-sequence research efforts during their lifetime, we should be able to see that essentially no one will be able to get further back than five full generations in a lifetime of work, meaning no family can ever get back more than five full generations on their own. So let's assume that person puts in their work on five generations of ancestors and then passes away. The next generation in that family might decide to do the same thing, and they start over from where they are, and, even if they go back a full five generations, the distance back they will get in the records will be one generation less than the prior researcher might have accomplished. So this means that we are actually losing ground in most cases when we hand off the research to the next generation.

Now let's suppose that there is one active genealogist for each extended family. You might imagine that if we had this extended family working on this project for 32 generations, with one active genealogist working during each generation, then you might reach the end of the 10 generations. However, most families are not able to be that persistent over a 400 year period, so that they could actually eventually finish the 10 generations. And, of course, if they wanted to go back 15 generations, that requires 32 times more effort, meaning that it requires 1024 lifetimes to complete, starting at any particular beginning person. We can be fairly sure that no family will be able to start a process which lasts the necessary 1,0240,000 hours of work or 10,024 years of work. (10,000 hours or 10 years times 1024 which equals either 1,0240,000 hours of work or 10,024 years of work.)

In other words, if we have any desire and hope whatsoever to finish the nation's and world's genealogy, we will need to find a way which is at least 1000 times faster than what we are doing today. Otherwise we will just wander about aimlessly and never finish any significant portion of it. Generation after generation will pass, and ancient records will be lost, and we will never make much more progress than we have already. On the other hand, if we get organized, and do things at least 1000 times more efficiently, we can get done fairly quickly and make sure that all the ancient records are used and incorporated before they become extinct.

2. "But we are sure that all of the research which everyone else did would be very bad and would be useless to us, and so we would have to do the work again ourselves, which is what we love to do, so there is no reason to cooperate."

This argument is likely to come from someone who feels confident that they can do a high-quality job themselves on the research, but who is sure that the other 4 million people in the country will do a really terrible job, so they will have to do it all again themselves anyway.

It is true that to use this system, one would have to have some confidence in the quality of research being done by other people. But it seems possible that the tens of thousands of people who belong to genealogy organizations and attend genealogy conferences might actually learn something from those experiences and be able to do high quality work on the data which they have practically sitting in their front yard, and which they are likely to be very familiar with already. Perhaps if they join in the project they will realize that the quality of their work is going to be examined potentially by many people, so they will feel a need to do a good and defensible job. They will realize that if they don't do a good job, they can't expect anyone else to do a good job either, and they will suffer from getting bad data from others.

Notice that the amount of duplicate research should be very minimal, since, by definition, people will be doing work on the data in the areas where they live (or on just the people with their surname, usually in quite a restricted geographical area.). If their research techniques draw them outside their area (or outside their surname), then they can probably guess that they are "doing someone else's work, which may be duplicated" and are not being very efficient.

3. "Cooperating with other people is just too boring and too much hard work so I don't think it's worth my trouble. The data I might get is probably of very low quality anyway, so why would I bother? Even when I do my part, hardly anyone else ever does there's, so it is just not worth it. I can easily put in a great deal more work than I ever get back in benefits from other people's work."

This objection has certainly been true in the past, but it would not be valid using the new software and methods I have developed. In the past, using old methods, almost any kind of "cooperation" meant tracking down people, many of whom might have moved, or stopped working on genealogy, or died, and sending out e-mails to try to make contact, and sometimes composing long and complex answers to e-mails, etc., etc. This is kind of a random "cold call" kind of situation where one can spend an enormous amount of effort and get very little in return. Or if someone has sent you an e-mail and you want to answer it, it can take an enormous amount of effort to winnow through your own data to find what you have and then put it together in a form which you can transfer to someone else. This is a very labor-intensive and tedious task. And then the people receiving the data may claim it was their work, or make changes which degrade the value of the data and hurt your reputation, etc., etc.

Under the new rules in the system I propose, there should be no reason to do any e-mails at all, except in a few rare situations. The database presents all the best data of everyone who is working, and you can simply look online and find what they have done. There's no reason to try to track somebody down and see if they're still alive and still active in genealogy work, etc., etc.
If their best data is online, you simply look and see what they have, and decide whether it meets your needs or not. There's no composing of e-mails and no answering of e-mails.

Assuming you have chosen to use the same-surname technique, most of the people you would communicate with would be your same-surname cousins, and most of their communication would be done by simply adding data to the agreed-upon structure. Internal messaging might be useful to clarify any work they have done, but otherwise the whole process is mostly self-organizing.

4. "I don't want to cooperate, because I am fiercely loyal to my own family, and I don't want anyone else messing with my data or presuming to offer the data they have prepared."

That stubborn independent streak has its value, but if a family EVER wants to get back more than three to five generations of quality genealogy work, they're going to have to swallow their pride a little bit and cooperate with other people, all the while encouraging everyone involved to do the best possible job and be respectful of everyone else's relatives. Otherwise, we have already done all the genealogy work that can be done, at least on the first 50 million people who lived in the United States. Anyone who is not already done, will never get done.

Maybe this little overview will make more sense than my earlier more math-oriented presentations (although mathematics is always very useful to calculate the consequences of any particular technique chosen). A version of this process could work almost anywhere, so that, as soon as the United States was finished, we could move on to Europe.

If we change the examples above so that we have only Church members during the work, instead of a larger segment of US genealogists, it doesn't change things very much, except to slow down the process of little bit. If, for example, we had 300,000 Church member genealogists working on this project, that would mean that they would each have to complete (300 million/300,000 = 1000) names. That 1000 names would take them 10 times as long to do the job as if we had all US genealogists involved, but it might still be possible to finish the job within six months time, still allowing about one hour per name, which should be extremely generous. They might finish the job in three months if it only required one half hour per name. Or, if we changed the goal of the project a little bit, and limited ourselves to the 70 million people who died in the United States before 1930, that would be (70 million/300,000 = 233) names each, a task which might be finished in a single month. So, if we allowed them a luxurious time period of six months, they could easily finish it.

There are numerous other ways this process could be organized, with the first criteria being that the process is divided up into small units which everyone can quickly understand, so that researchers can specialize and become experts in a small part of the problem, and so that every public record can be handled only once, and all those names linked together.