CAN A LAYERED APPROACH TO HISTORICAL FAMILY RECONSTITUTION MAKE A DIFFERENCE IN THE FINAL MERGING RESULTS?

David S. Barss and Jennifer L. Kerns
Historical Family Reconstitution
Family History Department

Abstract

Family reconstitution projects are becoming a popular method to study populations in a specific region. These projects are useful not only to the family historian, but to scholars, including demographers, cultural historians, and social historians. As the popularity and relevance increases, it becomes necessary to define and refine reconstitution processes. This paper discusses and compares two possible methodologies that could be applied to a project to achieve reconstitution success.

Definition

Historical Family Reconstitution for purposes of this discussion is defined as the process of bringing together automated historical records to identify families and extended lineages, usually based on a specific geographic region or community.

Big Bucket Approach

This approach to family reconstitution takes the data sets (automated historical records) and uses record linkage technology to identify matching records across all of the resources, merging the matching records to create families and extended pedigrees.

Layered Approach

This approach suggests that there is significance in the order by which these data sets are brought together that will improve the accuracy of the final results. It uses the same record linkage technology to match and merge the records, but it works with specific sets of resources in a specific order, rather than all of them at the same time.

Defining the Order

Mimicking the general process used to do family history research identifies both the sources to be used, and the order by which they should be combined. In doing family history research we work from the present to the past. We start by identifying what we know about our family. We try to benefit from the work done by others as we use compiled local sources such as county histories and genealogies. Then we move to sources that will help us to further identify our family and locate where they are living, like census, land, and probate records. We conclude by using primary source records to add to, and document, what we have gathered from the previous sources. Primary sources could include church records, civil registrations, and cemetery or tombstone records, to name a few.

The family history research process can therefore be summarized to these guidelines:

- Work from the present to the past
- Benefit from the work done by others
- Continue with family identifying sources
- Conclude with primary sources to add to and verify what has been found

This generalized family history research process suggests an approach to historical family reconstitution that is perhaps broader than most efforts are currently using. It suggests groups of records that can be worked with independently, before they are brought together in a specific order, as layers of data. As each new layer is added to the most complete data set, it enhances the existing data and adds new information not previously included, thereby increasing the opportunities for subsequent layers to match and merge with the growing collection.

The layered approach to historical family reconstitution then becomes:

- Start with compiled lineage linked sources
- Continue with family identifying sources
- Conclude with primary source records

With the following guidelines:

- Work from the present to the past
- Merge within each data layer before combining it with previous data layers

Big Bucket vs. Layered Approach

In either approach, we undergo the process of gathering the data, automating or converting it to a common format, then cleaning it up and preparing it for merging. Using the Layered Approach, we would retain the different data sets and bring them together in a specific order.

With the Big Bucket approach, when the data is ready, merging is done across the whole database. Therefore, using the Big Bucket approach requires fewer steps and would appear to be a faster way to proceed. So we ask the question: ***Can a layered approach to merging the data improve the accuracy of our final results?***

Answering the Question

We used records from Norway to answer our question, consisting of automated data from the Clerical District of Sør-Fron in the county of Oppland, Norway. Our data collection for this locality included the following records:

Compiled Sources:	
Bygdebok (Farm Histories with genealogies)	3,777 individuals
Family Identifying Sources:	
1865 Norway Census – Sør-Fron	3,825 individuals
1900 Norway Census – Sør-Fron	3,161 individuals
Primary Source Records:	
International Genealogical Index (IGI) data for Birth, Christening, Marriage records – Sør-Fron	3,903 individuals
--	---------------------
TOTAL number of individuals in the data collection	14,666 individuals

We worked to reconstitute the Sør-Fron Clerical District using these four data sets in both the Big Bucket and the Layered Approach.

Testing Process

We used the desktop version of GenMerge by Pleiades Software, Inc. as our matching and merging tool. We chose GenMerge for several reasons:

- Pleiades Software, Inc. was willing to work with us as needed.
- GenMerge uses probabilistic record linking as a foundation.
- GenMerge works on a copy of the database, and does not destroy the original data.
- GenMerge allows a review of the merge sets, with an option to exclude bad matches.

All data sets were processed through GenMerge using the same settings. We also used Legacy Family Tree (Legacy) and Personal Ancestral File (PAF) to view and manipulate the data throughout the merging process.

For the Layered Approach we began by merging within **like data sets** to create the following data layers:

Sør-Fron Bygdebok	(merged within to remove duplicates from the data set)
Sør-Fron Census Data	(merged within & across 1865 + 1900 Norway Census)
Sør-Fron Vital Records	(merged within to remove duplicates from the data set)

The final layers were then combined by merging:

the Census Data **into** the Bygdebok Data, and then merging
the Vital Records Data **into** the combined Bygdebok/Census data set.

This resulted in a **reconstituted view** of the Sør-Fron Clerical District.

We carefully evaluated all of the matches created by GenMerge in this process. Those that were rejected were captured in “Exclude Files.” We combined the Exclude Files into one Big Bucket Exclude file to use on the Big Bucket Approach so that the match decisions used in each Approach would be the same. We used the Layered Approach Results as our true merged data set when we did our analysis of the results. We did not do any further manual review to find matches missed by the merging software or excluded in error by the reviewer.

For the Big Bucket Approach we put all four of the data sets into one database (the Big Bucket) and then we ran the merging software on that one file with the exclude file noted above. The results were again a **reconstituted view** of the Sør-Fron Clerical District.

Testing Analysis

We created a Microsoft Access database that contained all of the records in the Data Collection, with each record identified by the Data Set it came from, this represented our Master List. The PAF RIN numbers in this Master List were the same as those used in the merging files so that we could use Microsoft Access to quickly identify how well the two approaches performed.

Testing Results

The final results are as follows:

DATA SET	INDVS	MERGED	DIFF.	TRUE DIFF.
Full Data Collection	14,666			
Layered App. Results		11,391		
Big Bucket Results		11,505	114	
Less Big Bucket Errors (1,656) gives		9,849		1,542

Number of Merging passes to produce the final data set:
Layered Approach = 7
Big Bucket Approach = 1

Final Conclusions - ???? wins

Based on test results above, the answer to our question ***“Can a layered approach to merging the data improve the accuracy of our final results?”*** is: Yes it can. The Layered Approach is the winner. We did have to do more work, a total of 7 merging passes on smaller sets of data, verses one pass with the Big Bucket Approach, but we feel the accuracy of the final merged results is worth the extra steps.

We had anticipated that the Layered Approach would outperform the Big Bucket Approach, because it brings together the data sets that have the most data first, thereby improving the chances for matching. The improved foundation created by the first two layers (compiled lineage linked sources and family finding sources) would then make it easier for the final layer (vital records) to find a match to link with. It appears from the above results that this is the case.

Collaboration – Data Test Set Sharing

Experience tells us that no matter what endeavor we are involved in we usually don’t have all of the answers. We are very interested in feedback regarding record linkage methodologies, and in receiving information on other tools and processes for Historical Family Reconstitution that may be faster and/or more accurate, and eventually more automated.

We are also willing to share the Sør-Fron, Oppland, Norway data collection for use by others in testing. In addition, we have data sets for other communities that we anticipate being able to share once they are complete.

David S. Barss
barssds@ldschurch.org
801-240-1357

Jennifer Kerns
kernsjl@ldschurch.org
801-240-6840