A Prototype for Splitting Munged Persons in Family Tree

Randy Wilson, *wilsonr@familysearch.org*
Information Architect, FamilySearch

BYU Family History Technology Workshop, February 27, 2024

Introduction

In family tree databases, each person entry is typically meant to represent one real human. When two person entries represent the same human, they are typically merged into one person entry.

However, it sometimes happens that one person entry represents two or more real humans. This can happen in two ways:

1. through a **bad merge** of two person entries that don't really represent the same real human, or
2. **organically**, by taking names, events, relatives or sources about two or more different real humans and adding them to the same person entry.

Terminology. This paper uses the term "*person entry*" (or "*person*") to mean a person object in a database; "*real human*" (or "*human*") to mean the actual human who lived on the earth; and "*information*" as a generic term that includes names, gender, events (like birth, death, residence, etc.), attributes (like occupation), attached sources, relatives (parents, spouses, children) and even temple ordinances where relevant.

This paper uses the made-up term "***munged***" to describe a person entry that has information about two or more real humans, since this doesn't always come about because of a bad merge.

Examples. As an example of a bad merge, we might have a person entry Bob with parents Doug and Mary. We find another person Robert with spouse Judy and child Jimmy. We think it's the same real human, so we merge them, and then add a death date to Robert, only to later find out that they aren't really the same person after all. Now we have a person entry with one real human's parents, a different real human's spouse and child, and a death date that probably belongs to one or the other of them.

As an example of an organic munge, we might start with a person entry Bill with parents Dave and Maude. We find a birth record listing William and Sarah as the parents of Patrick, in the same small town. So, we add Sarah as the spouse and Patrick as the child of Bill, and change his name to the more complete "William." Then we eventually find out that Bill and William weren't really the same person after all.

In both cases, we have ended up with a single person entry that represents two different real humans. In the first case, the munge happened because of a bad merge between two person entries, and in the second, it happened organically as information was added to the person entry.

FamilySearch Family Tree. Almost any family tree system can have duplicates, requiring a way to merge persons. And any family tree system can get into a state where a person entry has information about two or more real humans. So, any tree can have munges.

FamilySearch's Family Tree is a unified tree built by millions of users. The great advantage of this approach is that users can pick up where others left off, thus avoiding the exponential

amount of duplicate research inherent in approaches where everyone has their own tree. And the main challenge of this approach is that users can "step on each others' toes." Bad merges and other work resulting in munged persons is one of the most difficult challenges to deal with, so having a way to split munged persons is important.

The Family Tree does have a merge "undo" feature that works fine. However, this can only be used until any other change has happened to the person (including the merge of one of their relatives), so it is often unavailable. The Family Tree also has a "restore" feature that will take a "duplicate" person from a previous merge and restore them to look like they did before. However, it currently fails to remove any of the information that was added to the "survivor" during the merge, leaving the user to manually figure out what to remove from the survivor and then manually do so. It also doesn't currently help the user to know what to do with information that has been added to the survivor since the time of the merge.

To be clear, this paper is not announcing any planned features for FamilySearch's Family Tree. It is addressing the general need to split munged persons in any tree system, and uses a prototype to illustrate some approaches, using data that happens to come from Family Tree. Hopefully Family Tree and other tree systems can benefit from the suggestions found here.

Fixing a munge. Fixing a "munged" person entry can be tricky. It requires:

a) analysis to figure out what information is correct about the multiple real humans represented by the munged person entry,
b) creating a new person with their correct information, and
c) removing incorrect information from the remaining person.

Doing this by hand can be extremely difficult. This paper suggests approaches that can help with the process, and links to a prototype that demonstrates some of the suggestions.

Analysis

One of the trickiest parts of fixing a munged person is the analysis it takes to understand who the real humans are, and what each should look like. There are at least three tools that can be used to aid in this analysis, each of which provides a separate insight into the data, and each of which may be helpful in some situations and not in others.

1. Grouping Attached Sources. When the munged person has attached sources, it is usually best to start there. The information on the attached sources can often be used to reconstruct what the two (or more) real humans looked like, and therefore what information should end up on the two resulting person entries after a split.

Each attached source "persona" is part of a structured record that can have personal information (names, gender, events) as well as relationships (parents, spouses, and children). By sorting the information in those attached sources in various ways, it is often possible to figure out some common traits that distinguish one real human from another. By moving such sources into a group, we can often paint a picture of the two real humans, and thus figure out what the two resulting person entries should look like.

For example, consider the case in Figure 1 taken from real data on one of my munged ancestors. You can see in there a couple of spouses, and a couple of different record places. There are also two appearances in the same 1910 Census, which is suspicious.

	Source	Record Year	Record Place	Name	Birth	Parents	Spouse given	Spouse surname	Children
1	1900 Census	1900	Bell, KY	Massie Ball	1900, KY	John & Martha Ball			
2	Kentucky Marriages	1906	Harlan, KY	Mossie Ball			R. B.	Skidmore	
3	1910 Census	1910	Harlan, KY	Mossy Skidmore	1889, KY		Robert B.	Skidmore	Carrie, b. 1907
4	1910 Census	1910	Bell, KY	Massie Wilson	1891, VA		Charlie	Wilson	Thelma, b, 1906; Sarah, b. 1908
5	Kentucky Deaths	1918	Harlan, KY	Mossie Ball			Robert	Skidmore	Robert Clay Skidmore, b. 1917
6	1920 Census	1920	Harlan, KY	Massie Skidmore	1892, KY		Bob	Skidmore	Carria, b. 1908; Juanitta, b. 1915
7	1930 Census	1930	Bell, KY	Masie Wilson	1891, VA		Charles C.	Wilson	Thelma, b. 1907; Sarah, b. 1909
8	Kentucky Deaths	1936	Bell, KY	Mossie Ball			Charles C	Wilson	
9	Find-a-grave	1984	Bell, KY	Mossie Ball Wilson	Oct 1890				

Figure 1. Summary of information in attached sources.

Sorting the above data by the spouse name clusters several lines together well. Sorting instead by "Record Place" does an even better job. We can see in Figure 2 that the two resulting groups each have just one appearance in the 1910 Census, and have consistent spouses and record places.

	Source	Record Year	Record Place	Name	Birth	Parents	Spouse given	Spouse surname	Children
1	1900 Census	1900	Bell, KY	Massie Ball	1900, KY	John & Martha Ball			
4	1910 Census	1910	Bell, KY	Massie Wilson	1891, VA		Charlie	Wilson	Thelma, b, 1906; Sarah, b. 1908
7	1930 Census	1930	Bell, KY	Masie Wilson	1891, VA		Charles C.	Wilson	Thelma, b. 1907; Sarah, b. 1909
8	Kentucky Deaths	1936	Bell, KY	Mossie Ball			Charles C.	Wilson	
9	Find-a-grave	1984	Bell, KY	Mossie Ball Wilson	Oct 1890				
2	Kentucky Marriages	1906	Harlan, KY	Mossie Ball			R. B.	Skidmore	
3	1910 Census	1910	Harlan, KY	Mossy Skidmore	1889, KY		Robert B.	Skidmore	Carrie, b. 1907
5	Kentucky Deaths	1918	Harlan, KY	Mossie Ball			Robert	Skidmore	Robert Clay Skidmore, b. 1917
6	1920 Census	1920	Harlan, KY	Massie Skidmore	1892, KY		Bob	Skidmore	Carria, b. 1908; Juanitta, b. 1915

Figure 2. Source information after sorting and grouping.

We can use the above grouping to automatically suggest what a new person should look like, namely:

- Use the name "Mossie Ball" (especially if that is the most recent name before this person was merged).
- Move the bottom four sources over to the new person, removing them from the old one.
- Move the couple relationship with Robert Skidmore over to the new person, especially if the R.B. / Robert B. / Robert or Bob Skidmore in those four sources are also attached to the corresponding spouse in the tree.
- Move the parent-child relationships to the listed children over to the new person.

Note that in some cases, the latest conclusions on the current person may not include conclusions that should be chosen for one of the resulting persons. For example, if "1891, Virginia" is the latest birth date and place on the current person, then the split-out person might need to dig down into earlier values from the hierarchical change history, or into values on the sources, to provide one suitable for the new, split-out person, like "1889, Kentucky."

The prototype introduced below has a "Source View" that allows for the examination, sorting and grouping of attached sources, similar to the table shown above.

2. Original Identities. New person entries are usually created with a single real human in mind. Information added to a new person are usually—though not always—still talking about the

same human. But additions or merges that happen later may bring in information about a different real human, resulting in a "munged" identity. Being able to see what original identities were intended can provide a starting point for deciding how to split a munged person entry. This is especially helpful in cases where there aren't attached sources on a person.

Therefore, the prototype described below includes a "Flat View" that shows what the various merged persons looked like when first created (i.e., when migrated into Family Tree in 2012, or within 24 hours of creation). Any information that was added to the person after that time is shown with a "+" in front of it to indicate that it is an addition.

3. Merge History Hierarchy. If a munge happened because of a bad merge, then it can sometimes help to look at the hierarchy of merges that have happened through time. If we can see where the bad merge happened, and not too much has happened since that time, then sometimes we can select one "node" in the merge history hierarchy that contains under it all the information that needs to move out to the new "split" person.

For example, consider the following hierarchy of merges shown in Figure 3. The bottom (both with a spouse Robert Skidmore) were merged into the middle one, which was then merged into the 2nd from the top (with spouse Charles Wilson), resulting in the top node.

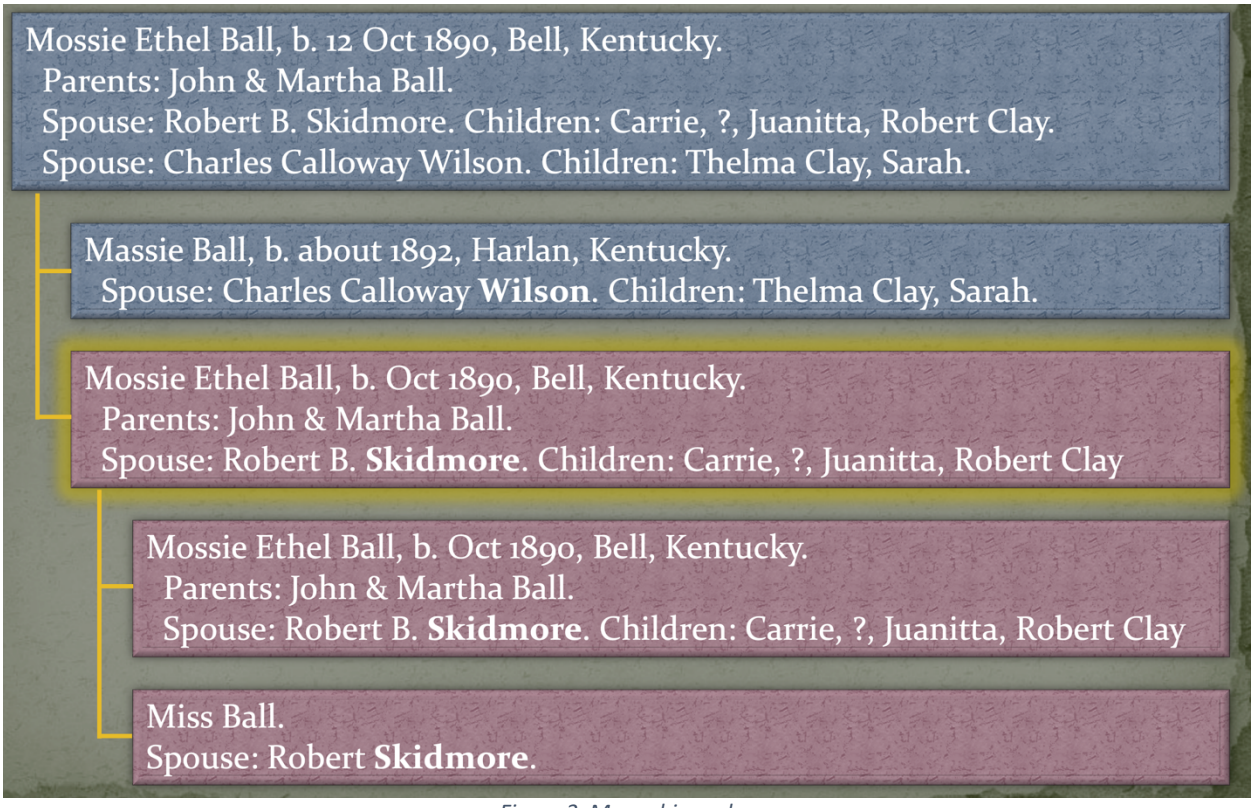

Figure 3. Merge hierarchy.

In looking at the sources and other data, you come to realize that the "Mossie" married to Robert is not the same real human as the "Massie" married to Charles. Each of these was fine up until that merge, but that merge was wrong. (And other data may have been added since that time). So, you could select either of those two nodes as the starting point for splitting out a person. The names, events, parents, sources, spouses and children on the selected node are added

to a new person (probably with the same old person ID as the one being pulled out), and any of that information that is no longer supported by the remaining person's remaining sources are removed from that person.

4. Manual splitting. There will be some cases where the above tools won't be sufficient, like when there are no sources attached and no merges have been done, and yet we still have a combined identity. In this case, a user will still need a way to split the person manually, after whatever research they can to arrive at their best attempt at the truth.

Usually, though, the above approaches will help greatly with the analysis of a munged person to determine which information belongs to each of the real human involved. Even then, however, once sorting and grouping is done using one or more of these approaches, the user will need an opportunity to review what the split person will look like and make adjustments before proceeding with an actual split.

Split Prototype

The above approaches seemed like they would be very helpful in fixing munged persons. However, "few designs survive first contact with the data." To test the usefulness of the above approaches and refine each to overcome problems, a prototype was created. By using the prototype on various real-world cases, some obvious shortcomings were discovered and addressed. There is still much more to learn.

The prototype reads data from the Family Tree public API. It does *not* write anything to Family Tree, so it is currently only meant to help with the analysis of a munged person to explore what a useful split tool might look like.

You can try it out by logging in to familysearch.org and then going to:

https://tinyurl.com/split-prototype

which will redirect to:

https://www.familysearch.org/service/gen/sforge/time-machine.html?pid=LZBY-X8J

Replace the person ID (pid) at the end with a person from Family Tree that you suspect is munged. (Please don't modify the person pointed to—it is already just one real human, despite the given name being written incorrectly as "Andrew" in a child's marriage record).

Click the "Help" tab for specific, up-to-date instructions on how the tabs work. Below are some principles illustrated in the prototype.

There are several tabs, which help you explore the data in various ways.

1. **Change Logs**. This shows all the entries for the change log for each person ID that was ultimately merged into the final person. Hover over an entry to see its details. The columns are sorted by when they were merged into someone else, with the final survivor on the left. This is not meant to be especially helpful in fixing a munged person—it is just a few into the raw data being used to build views in the other tabs.
2. **Merge view**. This shows a hierarchy of merges. The "leaf nodes" are the "original identities".
3. **Flat view**. This shows the original "intended identities".
    a. If "show additions" is checked, then it also shows a "+" in front of information added after the initial creation (i.e., migration in 2012 or 24 hours after creation).
    b. If "include deletions" is checked, then it also shows information that was added and then later deleted (in red, with strikethrough text).
    c. The "Facts" setting determines whether to show all events, just the "Vital" events (birth, christening, death, burial and marriage), or no events. Again, the point is to

allow you to zoom in and out of the data to find the right information to make decisions without swamping the screen.

4. **Sources view**. This view shows the list of attached sources, and what information each one contains.
    a. Click on a header to sort by that column (or on the "place" header to sort by place, highest level first).
    b. Click and drag the column headers to change their width.
    c. Click a row to select or deselect it. Shift-click will select a range.
    d. Click "New group" to move the currently-selected rows to a new group, or "Add to group" to add more selected rows to an existing group.
    e. Click on the name (like "Group 2") to edit the name of the group, if that is helpful.
    f. Click in a cell in the "Notes" column to add notes that might be helpful to you.
    g. All these controls also work in the "Flat" and "Combo" view.
    h. Click "Apply to Split" to use the contents of that group to update the elements in the final "Split view".
5. **Combo view**. This view includes all the Family Tree person rows from the Flat view, along with all the sources from the Sources view. Sorting by person ID groups attached sources under the Family Tree person that they were first attached to. One advantage of using this view is that you can decide which sources and Family Tree persons go into the group being split out, which helps get more of the information sorted to the correct side in the split view, when clicking "Apply to Split."
6. **Split view**. This view shows each item of information on the current person, with the ability to move the item between the original, "remaining" person (on the left), and the new, "split" person (on the right).
    a. Click on <, = or > to move that element to the left or right, or to copy it do both.
    b. Click on a "+" or "-" button to the right of any heading to reveal additional values that are not on the current person, but which were on a previous version of one of the merged persons, or are on an attached source. Check the box next to any of these hidden values to keep them visible when collapsing them using "-".
    c. Create a Group in the Flat, Sources or Combo view and click "Apply to Split" to initialize the Split view with suggested splits, using logic that takes the groups into account. Or select rows in the Merge view and click "Apply selected rows to Split" to do the same thing.

As stated above, the prototype does not actually split the person, though it might help you figure out what the person should look like, so you can do a "restore" and then clean up the two resulting persons.

Rather, the prototype is being used to explore how to make a user as efficient and effective as possible at correctly splitting a munged person. The prototype still has bugs and weaknesses (the relatives' information is their latest information, for example, rather than how they looked at the time that the original identities were created). But hopefully it will be helpful in exploring how to help users properly split person entries that represent multiple real humans.

If you have comments, questions or feedback, please send an e-mail to the author at wilsonr@familysearch.org.