Source Linker: Bridging the Evidence-Conclusion Gap

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

Abstract. *Family history research uses information in sources as evidence to draw conclusions in trees. While tools have been available to manage conclusional tree data for some time, there is a shift in the industry towards modeling the information in the sources themselves, as well as the links between persons in trees and their appearances in those sources. This paper reviews the data model behind this approach, and describes the FamilySearch "Source Linker" feature that has helped users take advantage of this approach to evidence-based family history research.*

Family history research involves using *information* found in *sources* as *evidence* to draw *conclusions* [1]. This was originally all done on paper, but in recent decades there have been many software tools to help researchers to organize their conclusions and cite their sources, including desktop family tree managers such as Personal Ancestral File (PAF). While these systems were helpful in modeling conclusions and citing sources, they did not directly model the information found in those sources.

The traditional data model used by PAF and other GEDCOM-compatible systems models a tree full of *persons*, each of which represents everything we have been able to learn about that real person so far from all of the sources available. The GenTech Genealogical Data Model (GDM) [2] introduced the concept of a *persona*, which can represent the information about a real person as found in an appearance in a *single source.* A real person can appear in many sources, and therefore more than one persona can be used to model those various appearances. (The GDM also uses a persona to represent groups of personas that are believed to be the same real person, with a conclusional person as a highest-level persona. In this paper, we will use *persona* in the more limited sense of the information about a real person that appears in a single source.)

In recent years, FamilySearch and several other organizations have modeled information from historical records and other sources using a persona-based model [3]. The idea is to model *what the sources say* about the people they mention separately from *what we conclude is true* about those people. So while *trees* hold information about what we believe to be true about a collection of real people, *record collections* hold information about what those records say about the people that they mention.

A core activity of genealogy has always been to determine which names in various sources are talking about the same real person. This is one of the most important genealogical conclusions we make, and is often the one that must precede all subsequent conclusions. For example, we must first decide whether this birth certificate is even talking about a particular ancestor before we move on to use it as evidence for that ancestor's birth date.

Once record collection data is modeled in a way that software can understand what records say about the people they mention, it becomes possible to link persons in trees to personas in records. This allows the system to know what each source says about the person in the tree. This, in turn, provides a variety of powerful advantages, including the ability to:

1. Tell the user when there is **additional information** in the source to add to the person in the tree;
2. Tell the user when there are **additional relatives** in the source that could be linked into existing relatives in the tree, or that could be added as new relatives.
3. Tell the user when there is **conflicting data** that the source could help resolve.
4. Use what the linked sources say to help users **decide on individual conclusions** (e.g., show what all the sources say about this person's birth date to help them decide the best one).
5. Use what the linked sources say to help users **split a badly merged person** in the tree (e.g., if one persona had these parents, and another had that spouse, and you decide those personas don't

represent the same real person, the system can often figure out where the relationships go on the resulting two persons in the tree).

6. Show which personas in each source are **already linked** into the tree, and which ones still need to be "accounted for." This helps many users avoid duplication of effort and thus work together to incorporate data from many sources into a shared tree "once" and "for all" [4].

While the paradigm of linking persons in a tree to personas in sources is powerful, it has taken some time for the ideas to catch on and appear in software. Fortunately, this is now starting to happen in various places. Ancestry.com, for example, has had their "shaky leaf" feature for some time, showing users personas from records that appear to match persons in their personal trees.

FamilySearch recently released a feature that is internally referred to as the "Source Linker." It shows data from a historical record on the left and data from the FamilySearch Family Tree on the right. The goals of this feature include helping the users to do the following.

1. First, decide if the person mentioned in the record is the same real person as the one in the tree.
2. If so, link ("attach") the persona in the record to the person in the tree.
3. Then, link any relatives in the record to corresponding relatives in the tree (moving them into the right position if necessary).
4. If there are new relatives in the record, add those new relatives as new persons in the tree.
5. For all people thus linked or added, if there is additional information in the record (e.g., a birth date), add that information into the tree as a conclusion.

In the example shown in Figure 1, "Martin L. Bumgardner" as found in the 1910 U.S. Census is attached to "Martain Luther Baumgardner" in the Family Tree. The spouse, Mary, appears in both the record and the tree, but has not yet been attached. The user can click "Attach", review the detailed information in the record and the tree, copy any new information over to the tree if available, and then finalize the attachment.

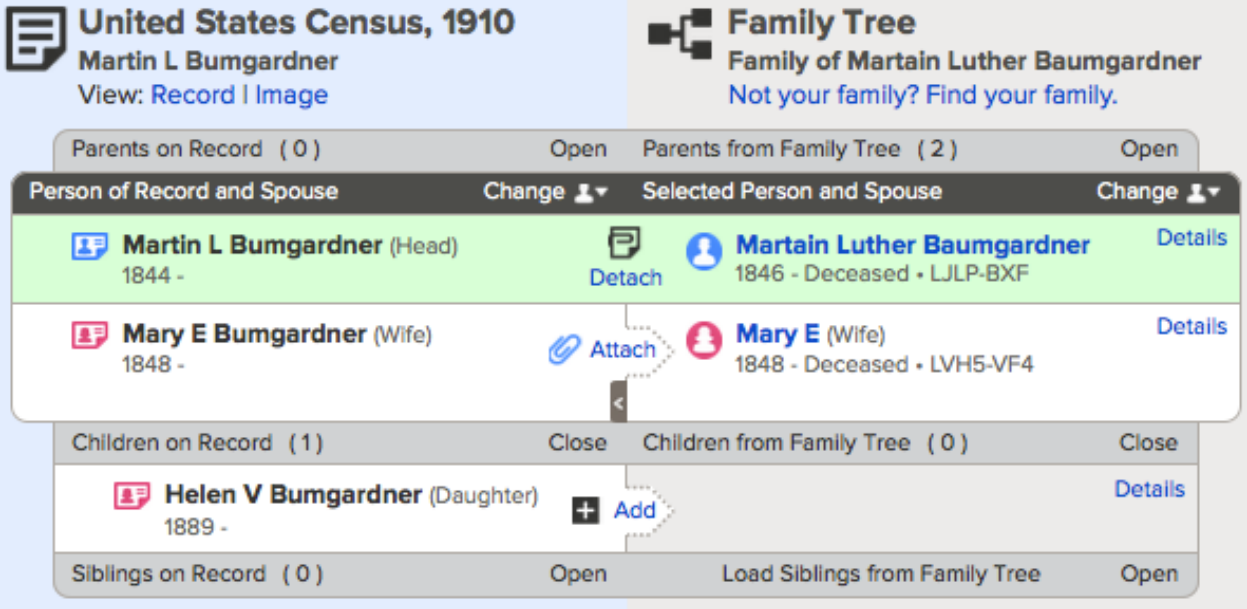


Figure 1. Martin L. Bumgardner in the 1910 U.S. Census is attached to Martain Luther Baumgardner in Family Tree. The wife Mary still needs to be attached, and the daughter Helen can be added to the tree.

The daughter, Helen, appears in the record, but not the tree, so she can be added from the record into the tree as a new child of these corresponding parents. Note that this is core to how genealogical research

is often done: we use sources that mention people we know about to discover new relatives we didn't know about, which helps us verify what we knew and add new people to the tree.

Once Mary is attached and Helen is added and attached, all three personas appear green on the left, indicating that the record is completely linked in, as shown in Figure 2. The next time a user visits this record, they can see that it is already linked in, thus avoiding potentially infinite duplication of effort.

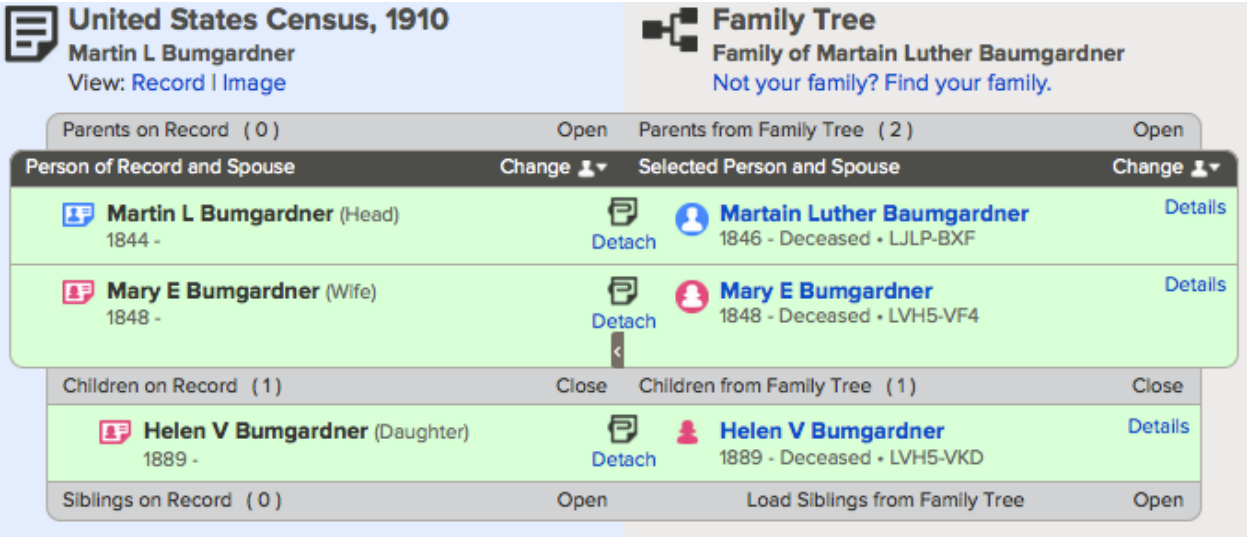
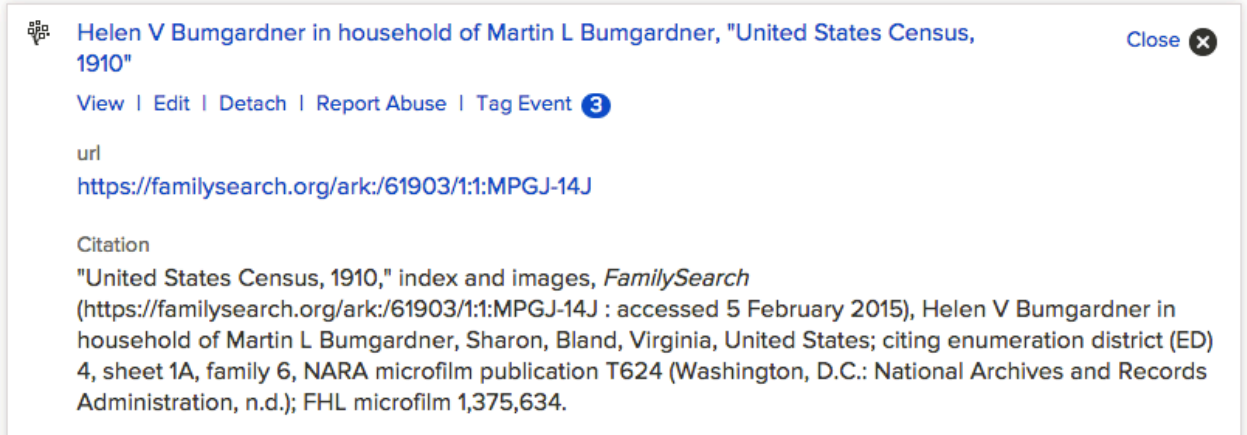


Figure 2. All 3 people in this record have been properly linked into Family Tree.

When looking at the list of sources for Helen V. Bumgardner in Family Tree, a source now appears like that shown below.



Note some important things that happened here.

- **Source citation.** There is a full source citation provided here, but the user never had to type it! The source "knew its own citation," and thus spared the user from the pain often associated with coming up with a quality human-readable source citation [1]. One expert can figure out a good source citation for all of the records in this collection, and then each user gets the citations for free.
- **URL.** There is a URL that a user can use to jump right back to the web page for this record. In fact, the URL happens to be a long-lived *Archival Resource Key (ark)* [5], so it is less likely to ever break. Having a URL to the source allows future users to see what the sources say with a single click. It also acts as a persona identifier that can be used for comparisons and reverse look-ups [4].

- **Structured record data.** The computer can use the same URL (with the right "Accept" header) to fetch computer-understandable record data in GedcomX format [3, 6]. This allows the software to support features such as those listed above (like telling the user what all the sources say about the birth date; indicating when there are additional relatives that aren't yet linked in; etc.)
- **Evidence-based conclusions.** The conclusions (e.g., names and relationships) added for this person came from genealogical source data, and the system (and future users) know where it came from. This makes it more likely that the conclusions are correct, and leaves less question as to where the information came from.

So, without having to be an expert genealogist, the user has done some good genealogical research and properly cited their sources in a way that both users and the software can understand.

FamilySearch has also recently introduced *record hinting*, which does matching between historical records and Family Tree data, and shows high confidence matches to users so that they can use the source linker to resolve matches, link relatives, and add new people. This combination of features has taken many novice users who were "stuck in the tree", and has helped them to become aware of available historical records that now help them to validate and grow the tree in a genealogically sound way.

Future directions

In the near future we will hopefully see more features that take advantage of the powerful model of personas in records linked to persons in trees. Features that help a user look at what the sources all say when deciding on a conclusion can help conclusions be more soundly based on real evidence. Tools to help users split a badly merged person using what the sources say can help them get the resulting people correct with much less guesswork and pain. Features that filter out already-linked records from search results will help users focus their attention on the records that still haven't been accounted for yet, which will avoid wasting time on records that have already been dealt with, which becomes more important as more and more records are linked into a tree.

The use of the GedcomX data model [3, 6] has made it so that each record persona is now an *entity* that has its own unique identifier. As more organizations adopt a common data model, it will hopefully become possible for tools such as the source linker to use data from different archives without having to write special code for each one. As more tools become available for a common data model, more services will benefit from supporting that model, and vice-versa, leading the industry to "explode with innovation."

The source linker is an important milestone. It bridges the evidence-conclusion gap and helps users build conclusions solidly on the foundation of information found in sources. As the industry continues to embrace the paradigm of modeling information in sources in addition to the conclusions drawn from it, family history research will continue to become more efficient and effective.

References

[1] Mills, Elizabeth Shown. *Evidence Explained: Citing History Sources from Artifacts to Cyberspace*. Revised edition. Baltimore: Genealogical Publishing Co., 2009.

[2] GenTech Lexicon Working Group: Anderson, Robert Charles, Paul Barkely, Robert Booth, Birdie Holsclaw, Robert Velke, John Vincent Wylie, (2000). *GenTech Genealogical Data Model, Phase 1: A Comprehensive Data Model for Genealogical Research and Analysis*. http://www.ngsgenealogy.org/cs/GenTech_Projects, 2000.

[3] Wilson, D. Randall, "Beyond GEDCOM: Modeling Genealogical Source Record Data." *Family History Technology Workshop 2013*. http://fht.byu.edu/archive/2013.

[4] Wilson, D. Randall, "Bidirectional Source Linking: Doing Genealogy 'Once' and 'For All'." *Family History Technology Workshop 2002*. http://fht.byu.edu/archive/2002.

[5] ARK (Archival Resource Key) Identifiers. *California Digital Library*. Accessed 4 Feb 2015. https://wiki.ucop.edu/display/Curation/ARK

[6] GedcomX open specification. http://gedcomx.org