High-Level View of a Source-Centric Genealogical Model: “The Model with Four Boxes”

January 25, 2006

Randy Wilson (wilsonr@ldschurch.org),
David Ouimette, Tom Creighton, Lynn Monson,
David Fox, Dan Lawyer, Steven Law, and John Heath

Abstract. This paper presents a high-level genealogical model intended to avoid infinite amounts of duplicate effort that is currently required in several parts of family history work. The model contains just four high-level elements: (i) a *source authority* keeps track of all known sources of data in the world, including written records and living memory; (ii) an *artifact archive* stores scanned images of documents and other digital artifacts; (iii) a *structured data archive* stores structured data that was extracted directly from sources; and (iv) a *family tree* stores conclusions about real people, including pointers to entries in the structured data archive that reference a person. In addition to tracking sources, evidence and conclusions, *verification* work is also tracked so that this, too, need not be repeated indefinitely.

1. Introduction

One of the ultimate goals of family history work is to create the most complete and accurate genealogical database possible from all of the sources that can be found in the world, and then do temple ordinances for all of these people.

The tasks required to do all of this can be summarized as follows:

1. Identify all of the sources of genealogical data in the world (including written records, living memory, and various artifacts).
2. Extract all of the relevant genealogical data from these sources into a structured digital format.
3. Link multiple references to the same person together, link related individuals together, and evaluate all the evidence to draw conclusions about who has lived and how they are related.
4. Verify that all of the above has been done accurately.
5. Perform temple ordinances for individuals and families.

There are several reasons why this could take an infinite amount of time given our current processes. While there has been much discussion on reducing duplication of temple ordinances, there is probably even more duplication going on in all of the other parts of the above work. It is often the case that spending time at a particular family history task does little or nothing to reduce the total amount of work that there is left to do. Often when one person does something, someone else will later have to do the same thing. For example, learning about sources in an area, flipping through an unindexed book looking for a relative, and spending time going back to verify conclusions drawn by others, are all activities that mostly benefit the person doing them, and will have to be repeated by anyone else who is looking for an ancestor in the same area or source, or who wants to verify the same conclusions.

This paper presents a high-level “source-centric” genealogical model, the details of which will be fleshed out as time goes on. This model is intended to begin very simply, in order to define just the elements necessary to make family history work possible in a finite amount of time. Additional details, enhancements, applications, tools and other extensions can (and should) certainly be added on top of this foundation.

2. High-level source-centric genealogy model

The model has four main elements:

1. *Source Authority*, which tracks all known sources in the world.
2. *Artifact Archive*, which stores images and other artifacts.
3. *Structured Data Archive*, which stores structured genealogical data extracted from sources.
4. *Family Tree*, which stores conclusions about who has lived and how they are related.

In addition, *verification* work is tracked, which is made possible by the existence of these elements.

Each of these elements is described in more detail below.

2.1. Source Authority

The source authority’s job is to maintain a list of all known sources of potential genealogical relevance in the world, and to assign each source a unique ID. It provides a way for a user (or library, etc.), to look up a source to see if it is already in the source authority, get its unique ID if needed, see what work has been done with that source, and see what remains to be done.

When looking in the source authority, a source could be in one of several possible states:

1. Not yet listed in the source authority.
2. Listed in the source authority (but not scanned or extracted).
3. Unique ID assigned to each page of the source.
4. Some or all pages scanned
5. Some or all pages extracted.
6. Some or all extraction work partially or fully verified.

By following links from the source to the structured data archive and on to the Family Tree, it would be possible to derive other states for a source as well, such as:

7. All extracted individuals from the source are linked into the Family Tree
8. All merging with that individual currently verified

Users and repositories (such as libraries, churches, and government agencies) would need a way to add new sources to the source authority, preferably with automated ways of detecting which sources are already in there.

The source authority would contain lists of sources such as census records, vital records books from county courthouses, will books, compiled family histories, land records, tax records, and other books. It could also support a variety of other types of sources, some of which may not immediately come to mind, such as:

1. Personal journals (even if unpublished)
2. Family Bibles (with family history information inside the cover)

3. The living memory of users (each user is a “source”, and each person they personally know or have been told about is a “persona” to which they can attach the information that they remember).
4. Electronic databases (some of which may be evolving over time), including personal record manager databases and organizations’ electronic collections.
5. Letters
6. Graveyards (e.g., someone with a digital camera could generate an image of each headstone [or groups of them], and then extraction could be done from these. The extraction project could be the “source”, and each image can be a “page”).
7. Encyclopedias or historical books (contain information about living people).
8. Any other book that references real people (even if it is only the author’s name and approximate time period and place. Note that in some of these cases scanning the pages would probably be overkill).
9. Pictures (especially with individuals identified within rectangular regions of the image)
10. Audio or video recordings (e.g., could transcribe what is said and extract data from that, or could even specify time within either stream where each extracted piece of evidence came from—or both, since aligning text to audio is the one thing that speech recognition software is actually really good at).

The source authority serves as the main “To Do” list of family history work. As sources are added to it, each source can then be tracked through its various states until it is extracted, linked into the Family Tree, verified, and all ordinances are done for any new people the source introduced.

Without a source authority, sources could be repeatedly extracted (in bulk or one person at a time) and linked into the Family Tree forever.

The source authority could also contain a list of all known repositories in order to know which repositories might still have additional sources that have not yet been added to the source authority. In addition, the source authority could optionally keep a list of which repositories had a copy of which sources. Keeping such a list up to date might be very difficult, however, so it might be left up to each interested repository to do, since the source authority’s job is to keep track of what sources there are and what has been done with them, and not necessarily to keep track of where every copy of each source currently resides. A *Repository Holdings Manager* could instead be spun off as a separate system that uses the source authority for information about each source.

2.2. Artifact Archive

The artifact archive’s job is to hold scanned images for each page of each source, and serve them up over the internet as needed for extraction and verification of extraction. It could also serve to permanently preserve a digital copy of each source. In addition it could hold other artifacts such as photographs of ancestors (with individuals later identified within rectangles), audio or video interviews (with portions later transcribed and extracted), etc., that can be pointed to from the structured data archive (and perhaps from the family tree in some cases).

Each image gets a unique ID for its source (and page, if relevant) from the source authority. The reason that the page ID should be assigned by the source authority rather than the artifact archive is that while the ideal would be to have a scanned page for all extracted data, some sources could be extracted without stored images. This might happen, for example, when (a) copyright laws would be violated by scanning the image but not by extracting the structured data; (b) a user wants to extract data from a source that is not slated for scanning any time soon [and the user is not able or is not allowed to scan it themselves]; (c) the source has very little genealogical data spread throughout many pages, so extracting straight from the book makes more sense [though scanning just the pages with relevant info might still make sense]. In this case, verification work would require getting physical access to the same source, but although this isn't as convenient, it could be done, and after several verifications, the extracted data could still be quite reliable.

2.3. Structured Data Archive

The structured data archive's job is to accurately represent the data in the source, as well as to provide a unique ID for each *persona* that appears in each source. A *persona* (a term and concept borrowed from GENTECH's Genealogical Data Model) is a reference to a person that appears in a source, and is not to be confused with a *person*, which is a real person. A real person can appear in many different records, as well as in the memory of living individuals. Each of these appearances constitutes a *persona* that is recorded in the structured data archive, and assertions can later be made to link such "personas" together into a person in the family tree.

It may well be that the user who is extracting a particular record knows additional information about the person that the record refers to. However, such additional information needs to be entered elsewhere. An extraction should be true to the source, and additions or even corrections (i.e., when data is wrong in the original source) should be made either by entering additional data from other sources (including the user's own knowledge), or in the family tree.

Some provision will need to be made in the structured data archive for handling redundant extraction (such as double-blind extractions; partial extractions followed by more complete ones; and extractions of sources that are derivatives of other sources). The structured data archive also has to be able to allow corrections or variant opinions when there are errors or disagreements in the extraction itself.

The structured data archive also needs to provide ways of searching its contents, both to find potential matches for individuals in the family tree, and from scratch.

In addition, it needs to provide a way of browsing extracted records in a way that takes advantage of the context available for each type of source. For example, browsing census entries such that you can see who is in next door is helpful, and browsing from a sorted index is useful as well.

2.4. Family Tree

The family tree's job is to represent the world's current best conclusions as to who has lived, what we know about them, and how they are related. This is where assertions are made about which "persona" records in the structured data archive are really the same real person, and other conclusions that can be drawn from the evidence. Data will likely be copied from the structured data archive for the sake of efficiency, and some sort of persona id will be used to point back to entries in the structured data archive.

2.5. Verification

With the above four pieces in place, it finally becomes possible to track verification work done by users. Genealogists are currently encouraged to go back to original sources to verify each conclusion they find in a compiled source—especially electronic ones. However, this implies that nobody is supposed to trust anyone else’s work, so only small parts of it can be declared “finished” by anyone, and it would never be possible to anyone to declare all of it as “finished”.

However, with the above model in place, we could store the verification “stamp of approval” of users on individual parts of the work. A user could agree that two “persona” records do in fact apply to the same person; that data about each persona was extracted correctly; that other conclusions were drawn correctly; etc. Conclusions could thus eventually attain enough confidence that other users could trust the verification work that has been done, and turn their efforts to other work that still needs to be done.

3. Conclusion

The high-level source-centric genealogical model outlined in this paper contains elements that are essential to avoiding infinite duplication of effort in family history work. Many tools and interfaces can be built on top of these elements to make family history work even more efficient and to provide good user interfaces, but these elements will likely form the foundation of a system that enables truly collaborative work.