Beyond GEDCOM: Modeling Genealogical Source Record Data

D. Randall Wilson, *wilsonr@familysearch.org*

Family History Technology Workshop, March 22, 2013

Abstract. *Computer software has modeled genealogical conclusions for decades using a data model similar to the one used by the GEDCOM data exchange format. Modeling genealogical source data that appears in historical records, however, presents some challenges beyond what a conclusional data model was designed to handle. This paper discusses the ways in which historical record data and conclusion data are different, and in what ways they are the same. It illustrates how a model can handle both in a way that makes them compatible for many uses while preserving some of the nuances found in the records.*

1. Introduction

Family history work was once done on paper. Original *source* documents were used to find *information* that could be used as *evidence* to draw *conclusions* [1]. Conclusions were then recorded on paper as well, such as on pedigree charts, family group sheets, or narratives. When done well, these conclusions were backed up by references to the original sources, *i.e.*, *source citations*, and sometimes even by excerpts from the sources and an explanation of the logic used to draw the conclusions.

Modeling conclusions. For several decades, conclusion data has been modeled on computers in desktop clients and large databases, and such data has been commonly exchanged via versions of the GEDCOM standard. Since the computer understood what the conclusions were, it could be helpful in printing reports, organizing information, visualizing the data, sharing data with others, and even identifying suspicious-looking conclusions that could use more attention.

Modeling record data separate from conclusions. More recently, the GenTech Genealogical Data Model [2] proposed the concept of a *persona*, a "reference to a person" as found either in a source document or in a conclusion tree. The idea of modeling "what the source said" separately from "what we conclude to be true" has opened up a new era of family history work.

It is said that the conclusion that "those two are both the same real person" is perhaps the most important genealogical conclusion we make. It is the foundation for most other conclusions, because it allows us to take information from one source and add it to what we already know about the person. To do this, it is essential that we model what the record says (so that there is even a persona to link to), and then model the link between the two.

With this paradigm, it is possible for software to understand what the sources say so it can be helpful in a variety of new ways, including the following.

- *Record match.* Suggest to a user what records seem to mention their ancestors
- *Data input.* Help a user bring information from a source directly into a conclusion tree
- *Suggest relatives.* Given a link from a person in a tree to at least one persona in a historical record, suggest relatives in the record that can be linked to corresponding relatives in the tree, or added as new people in the tree.
- *Resolve conflicts.* Identify cases where information in the source conflicts with the current conclusions, or where two different people in the tree link to the same persona in a record.
- *Reverse links.* Show which individuals in a record have already been linked into a tree.
- *Etc.*

By modeling *what the sources say* in addition to *what we conclude to be true*, and also the link between the two, a system can also help track what information in the source has been "accounted for" in a conclusion tree, and can thus help users avoid duplication of effort and focus on sources that have not yet already been fully dealt with.

Section 2 lists ways in which record data differs from conclusion data. Section 3 reviews things that are the same in both models. Section 4 provides a high-level model of how the two can be unified in a way the provides consistency for many uses while retaining the nuances of record data for others.

2. Unique requirements of record data

Difference in purpose. While "record data" and "tree data" have many similarities, there are some important differences as well. While tree data represents *what we think is true* about the people in a tree, record data represents *what we think the record says* about the people it mentions. Put another way, tree data contains conclusions about what we think is true about the people, while record data contains conclusions about what we think the record said or what it strongly implied.

If a conclusion tree contains an error, we correct the error by drawing a better conclusion. By contrast, if an original record contains incorrect information, we do not modify the record (i.e., we don't "alter the evidence to fit our conclusion"). Rather, we should account for the incorrect information when drawing conclusions in the tree (e.g., "I believe this person was born in 1901 even though that source says 1910. That was likely a typographical error, and she would have been 9 years old at the birth of her first child if that were correct….").

Difference in structure. Modeling data from source records has a few requirements that are not faced by conclusion "tree data". While a conclusion tree typically *normalizes* its data and puts it into a more consistent structure, record data often needs to capture nuances that are important to preserve. Some of the differences include the following.

Fields. Record data is often extracted from a form that has fields such as first name, middle name, surname, father's name, birth date, witness name, relationship to the head of household (e.g., in a census), "beneficiary's relationship to the deceased," and so on. Sometimes fields therefore have some semantic meaning that is not modeled in the conclusion world, but is helpful for users to be able to see. For example, knowing that someone is a beneficiary gives us a hint that they might be a spouse or child of the deceased, though we may not be able to go so far as to create a parent-child or couple relationship for them.

Rectangles. Fields may also have rectangles indicating where the fields were found on an image. This may even be true of fields in sources that do not come from a fixed set of fields, such as an obituary or will, where any number of people can be mentioned with arbitrary relationships described.

Original and interpreted values. We often take what a record originally said (e.g., "Wm." or "M") and interpret its meaning (e.g., "William" or "Male", respectively). However, it is possible that our interpretation is wrong on occasion (e.g., in a Mexican census, "M" means "Mujer" which means "Female"). It is therefore important to preserve both *what the record said* and *what we think it meant.* The original value can be thought of as the "source" of the interpreted value, and needs to be available for verification and possibly for correction of our interpretation.

Lack of augmentation. In a conclusion tree, we incorporate whatever information we can find out into the tree. For record data, though, while it is appropriate to include what the source says and what we think it meant, it is perhaps best to resist the temptation to also include information from external sources. For example, if a census lists someone's name as "C. B. Wilson" and I happen to believe that this is my grandfather whose name was "Collie Bryan Wilson", I might be tempted to introduce that information into the census record data. But the record didn't actually say that nor even imply it, and if it turns out that this isn't really the right person, then that data doesn't belong there. Such conclusions are better made in a conclusion tree with a link between the two persons indicating the belief that they both refer to the same real person.

Parts. In original records, sometimes data comes in full fields (such as a full date or full place), but sometimes it comes in *parts*, such as when a name is pulled from *first*, *middle* and *last name* fields; or when a date comes as *day*, *month* and *year*; or when a place comes as *county* and *city* (with an implied *state*). In a conclusion system, on the other hand, we tend to take those parts and draw a conclusion from them such as the full date, as well as format them to our liking. If we find data in parts in a record, however, and the type of the parts are identified, then we don't want to lose that information, as it can help us interpret the data correctly, especially where it isn't completely obvious (For example, is "6-7-1923" June 7 or July 6? If we know that "day=6" and "month=7" then we have removed the ambiguity).

Roles. Sometimes people mentioned in a record have a *role*, such as *principal*, *bride's father*, *witness*, or *beneficiary*. From some roles it is possible to infer parent-child or couple relationships. From others there comes simply an understanding of why the person was listed in the record, which can be useful in a variety of subtle ways. For example, the witness of a will might be a relative or neighbor.

Record-level information. Record data can also contain information that does not pertain to any single person in the record, but applies to the record as a whole, such as a document number, record type, etc.

Person grouping. A record is generally self-contained, meaning that all of the people in a record typically only have relationships to other people in the same record. It is also usually small enough to be retrieved as one bundle. Therefore, the simple fact that two people appear in the same record gives them an affinity that can be valuable, such as when supporting queries such as "find me people named Robert Thompson in the same record as someone named Zedekiah."

Collections. Records are very often grouped together into *collections*, which logically group resources that have a similar origin or structure. Often there is a natural order to the records in a collection that reflect how they were created or organized, which can in turn provide powerful context.

3. Similarities of tree and record data

In spite of the differences in how tree and record data need to be treated some of the time, in many ways they are the same, in that both types of data include information about people, including names (with given name and surname importantly identified); gender; events (with types, dates and places); and relationships to other people. In fact, sometimes one tree is used as the "source" for building another tree, so it has turned out to be extremely convenient to be able to access information about people from either a tree or from records in the same way much of the time, without an unnecessary mapping exercise to get in the way.

When displaying detailed information about a record, it is helpful to be able to see the nuances discussed above. But when determining what information a source contains that might be helpful to make use of in a conclusion tree, it no longer matters whether that information came from a record or from some other conclusion tree—the structure of that information is basically the same, as far as the tree cares about it.

4. Reconciling the models

Given the above discussion, this section provides an example of a model that allows information about persons to be modeled in a single way, with additional record information added in.

Conclusion tree. At a high level, a conclusion tree can be defined using the following concepts.

- A *Tree* is a *Collection* of *Persons* and *Relationships*.
- A *Person* has *Names*, *Facts* (including events and characteristics), a *Gender*, and *Source Links*.
- A *Name* has *Name Forms* (each of which can be of a certain type or script, to support cases such as Korean, in which there may be a Hanja, Hangul and Romanized version of the same name).
- A *Name Form* has *Name Parts* with a *Name Part Type* (*e.g.*, given, surname, prefix, etc.) and a *Value*.
- A *Fact* has a type (e.g., birth, death, occupation, etc.), and can have a *Date* and a *Place*, each of which can have a *Value* (i.e., text) and a *Formal Value* (*e.g.*, a URI that uniquely identifies a place, or a date string that conforms to a standard).
- A *Relationship* has two person references and a *Relationship Type*, the most common of which are parent-child and couple. (From these, a Family can be derived to support traditional views with father, mother and children, as well as pedigree views, etc.).
- A *Source Link* has a URI defining the resource that is being pointed at and a *Source Link Type*, which indicates what meaning to give to the link. The most powerful kind is a *Person Link*, which is interpreted as a conclusion that "the person being pointed at in the source is the same real person as this person in this tree."

- Names, Facts, Relationships, Source Links and even the creation of a Person are all *Conclusions* that record who made the conclusion, when, how confident they were, and their reasoning.

Record collection. A *Record Collection* can be defined at a high level re-using all of the above concepts except for Tree, and with the following additions.

- A *Collection* is a group of related *Records* or other *Collections*, thus allowing a sub-collection hierarchy. (A Tree can be thought of as a special kind of Collection that has Persons in it instead of Records).
- A *Record* has *Fields*, as well as Persons and Relationships as defined above.
- A *Field* has an *original* value (indicating what we believe the record said) and an *interpreted* value (indicating what it "meant"), either or both of which could be text (human-readable) and/or a formal URI (computer-understandable). A field can also point at other fields from which it was derived (*e.g.*, if a full place field was derived from parts), and can have source qualifiers indicating, for example, what rectangles on an image a field came from.

One other important aspect is to know how the values in the Fields relate to the values in the Person and Relationship conclusions. Each Value can reference a Field Value to show where it came from. If a Field Value is modified, software can then know that a corresponding conclusion should be considered for update (ideally with a user's input).

By re-using the same elements to represent the same kind of information in records as are used in trees, software systems can avoid the expense, complexity and hassle of mapping between models constantly. For example, instead of having "record persons" with "record facts" with "record dates" that all have to be mapped, we can just have persons, facts and dates, which can generally be used interchangeably. Those operating exclusively with conclusion tree systems simply need to ignore the field value references that might seem superfluous in that context. This is a trade-off, but a small price to pay for the benefits of interoperable record and tree data.

5. Conclusion

Historical record data has some nuances and additional kinds of information that is not normally part of genealogical tree data. It is important to model and preserve such information so that important genealogical data is preserved. However, in most ways, a record is telling us the same sorts of things about the people it mentions as a tree tells us about the people it contains. A record therefore acts as a "mini-tree" for many purposes.

By modeling the persons and relationships the same in both records and trees, software can deal with either one in the same way for many purposes, such as drawing persons with their relationships; helping to bring new information into a tree; searching or matching on data; showing what all the linked source people say about a conclusion person in the tree; and so on.

Meanwhile, by having a list of fields and grouping persons by record, it is possible to preserve the additional richness that a record provides.

While conclusional models of genealogical data such as GEDCOM have been a great help to the genealogical community, the industry now has the data and the knowledge to make it possible to go beyond just tree data and model the source data upon which conclusions in the trees are based.

As systems are developed that model what the sources say, what users have concluded about the people, and how the two are related, we will be able to be far more effective in doing family history work that is soundly based on evidence found in sources, while avoiding unending duplication of effort.

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.