In *Proceedings of the 2nd Annual Family History and Technology Workshop*, pp. 54-60, April 4, 2002.
[Latest version and any updates available at http://axon.cs.byu.edu/~randy/gen/Bilink.html].

Bidirectional Source Linking: Doing Genealogy "Once" and "For All"

D. Randall Wilson
fonix Corporation
WilsonR@fonix.com

Abstract. There are many people working on genealogy throughout the world. Unfortunately, much of their time is spent duplicating the work of others or doing work that will be done again later. This paper discusses the reasons behind this duplication of efforts. It then proposes the use of bidirectional source linking in conjunction with evidence-based genealogical research methods to make it possible to quickly find out what work has already been done with regards to a particular record or set of records, and to more easily know what else remains to be done. The proposed methods also make it possible for genealogists to do work just once that can then benefit everyone.

1. Introduction

Suppose our goal was to build the most complete genealogical database possible from all of the records on Earth. It may seem like this would take forever, and that we would never be able to tell if we were done. In fact, using traditional approaches to genealogy, that just might be true. Millions of people work on genealogy, but a large percentage of their time is wasted on work that has already been done, or that will have to be done again.

To illustrate how efforts are duplicated and why this happens, consider the simpler (though still massive) task of using 100,000 volunteers to transcribe the first one million microfilms in the archives of The Church of Jesus Christ of Latter-Day Saints.

If each volunteer is given a specific list of rolls of microfilm to transcribe, and the transcriptions are collected into a centralized archive, then (after a few years, perhaps):

- Each record would be transcribed exactly once (or twice, perhaps, for verification), and
- We would know when the project is complete, because every microfilm would have a corresponding transcription (and vice-versa).

When the project was complete, people could do fast searches over the information, copy & paste relevant portions for their own use, etc.

However, consider what would happen if the volunteers did not know which microfilms were already being transcribed, nor where to put the transcriptions.

- Some records would be transcribed many times, and others not at all.
- There would be no way to know when the project is "done".
- Volunteers would have their own transcriptions on their own hard drives, but nobody else would benefit from their work, except people who are in direct contact with the volunteers or happen to find their transcription posted at a website somewhere. Furthermore, it would take so long to see if the transcription existed and locate it that it might be faster to just re-transcribe it anyway.

As ludicrous as that sounds, that is similar to how genealogy is currently being done. People search for records with references to their relatives. They may transcribe these references and extract genealogical conclusions from this evidence. They enter these conclusions into a genealogy database, hopefully with source information (though usually not more detailed than down to the page number). They share this data with family members, on a personal web site, or through one of several on-line repositories (e.g., *FamilySearch.org*, *Ancestry.com*, etc.).

If someone else is interested in the same individual and doesn't happen to find the electronic data on the web, then they go through the whole process again. Even if they do find it, they often need to go back and check the record to see what it really says, especially if it conflicts with previous conclusions.

Given a particular record, it would take a long time to see if all of that information has been assimilated into a particular repository—perhaps longer than re-entering it. This makes it highly likely that someone will eventually re-enter the information all over again. Furthermore, as

people gather copies of information they find in various sources, they add to the volume of records available, often without actually adding any new information.

This is not to say that there is no value in gathering information for one's own use. On the contrary, getting to know one's ancestors is an important reason for doing genealogy in and of itself. By avoiding the duplication of effort that is so common, more time could be spent reading about one's ancestors and enjoying the fruits of the research instead of spending so much time wading through unrelated information looking for bits of relevant information.

2. Evidence-based research

Genealogists should keep in mind that there is a difference between *evidence* and *conclusions*. For example, if a record says "John Smith was born on July 12, 1835", this is not necessarily a fact. The *fact* is only that the record *says* that.* This in turn is *evidence* that supports the conclusion there was someone named John Smith who was born then. If, in addition, a census record in 1850 lists a John Smith at 15 years of age, it is not a fact that this is the same person. Rather, the same name and the matching age and place are all evidence that could lead one to the conclusion that they are the same person.

When attempting to merge genealogical data, there are often conflicting conclusions. Two genealogists would usually agree on what the original sources actually say, but may disagree on the conclusions drawn from the evidence, especially if they have different evidence available to them.

Unfortunately, most genealogy software currently encourages users to enter conclusions first, and then optionally add source documentation.

Some people [1] [2] have recommended the opposite approach: Enter the evidence first, then make assertions from this evidence, and further assertions on top of these, using "preliminary conclusions to build more advanced conclusions." [1]

By keeping track of each piece of evidence and the logic used to make each assertion, every conclusion can be traced back to exactly where it came from. Conflicting conclusions can be traced back to where their assertions first diverged from where they agreed on the lower-level evidence and assertions.

The GENTECH Lexicon Working Group has created a Genealogical Data Model [1] in which they show what form the evidence, assertions and conclusions could take. They stress that software should encourage (and teach) the user to store "the reasoning behind the genealogical conclusions reached, along with all the evidence that led to those conclusions." When merging data, the low-level evidence and the logic behind each assertion up to the high-level conclusions are all shared. All data keeps an audit trail to show where it came from.

They also suggest that there should be exactly one place for every piece of genealogical information. You should not, for example, transcribe information, then paste it into a note and then type it again into a field in the database. This results in multiple copies of the same information which allows for errors and inconsistency. Rather, the database should point to where the information actually came from, and perhaps retain a local "cached" copy of the data for convenience (which could be refreshed if needed).

3. Bidirectional Source Linking

As helpful as it would be to store the evidence that leads to every conclusion, that still does not solve the problem of duplicating efforts and never being sure when a record has been "done." Even if someone meticulously keeps track of all of the evidence they used and the assertions made at each level, the next person that comes along and finds the same record would have a hard time knowing that anyone had ever looked at that record before, and they would be likely to repeat the work all over again.

---

* This paper is a good illustration of this principle, because it may very well be that there *wasn't* a John Smith that was born July 12, 1835, even though this document says so, because I totally made that up.

This paper proposes the use of *bidirectional source linking*. If such links are used properly, not only would every conclusion show what evidence led to it, but every piece of evidence would point to the conclusions that were derived from it.

When looking at a record that has been assimilated into a genealogical database using bidirectional source links, one could follow the trail to see what low- and high-level conclusions made use of each piece of evidence. By traversing the bidirectional evidence graph back towards other evidence, one could also see what other records were also used to assist in those conclusions. This could, for example, allow a genealogist to find out what other records have already been found that were believed to refer to the same person.

By continuing to follow the evidence graph, one could also follow relationship conclusions to find additional relatives of the referenced person. From any of the connected high-level conclusions, one could follow the links back to the original evidence used for those conclusions as well.

Under such a scheme, genealogists would not have to duplicate the effort of extracting and linking information, but could simply make use of (and perhaps spot-check) the work already done, and then, when a point is reached where the work has not yet been done, the genealogist could pick up where someone else left off.

4. Doing genealogy “once” and “for all”

By using bidirectional source links and evidence-based genealogical databases, genealogists may be able to finally avoid spending so much of their time duplicating the work of others. There are many ways in which such ideas could be put into practice, and it is not yet clear what approach would be best.

This section presents one example of an approach for using evidence-based genealogy with bidirectional source linking. Keep in mind that there are many details that would need to be worked out to make such a system practical. This set of steps is provided merely as an illustration of how these ideas might be put to use. Hopefully this example can foster further improvements by others.

Find a record. A genealogist would begin by somehow locating a record with some genealogical information in it. This may be either in a search for their own relatives, or may be a record that they have volunteered to “assimilate”.

Add the record to the master list. If the record is not already part of the master library, then the person “enters it into evidence.” This involves getting a unique, permanent identifier for it, and then adding its description and identifier to the master library so that it can be easily found in the future. This step only involves identifying the record (e.g., book, document, etc.), so it is possible for a record to be in the master list without any of its contents yet being in the master database. In fact, it might be common for many records to be added to the master list at once without any further processing done to it until later, e.g., when a library adds its entire collection to the master list.

Scan the record. If the record has not been scanned, the person could scan it and add the image to the master library. This allows the document to be preserved, as well as viewed over the internet for the sake of verification or for transcription and subsequent steps in the process.

Transcribe the record. If the record has not been transcribed, the person could transcribe it and add the text file to the master library. This allows for faster searching, indexing, efficient viewing over the internet, and further processing. Optical character recognition (OCR) might be used to speed up the process. It might also be useful to store not only the text, but also the rectangular coordinates of where each word (or line) appears in the scanned image so that that area can be given special attention upon verification of any extracted text.

Extract structure. If the record's text file has not had its basic genealogical structure extracted, the person can do so and add the structural information to the master library. For example, software might allow the user to highlight portions of the text transcriptions and type keywords, click buttons, or even use speech recognition to identify what kind of information each portion contains.

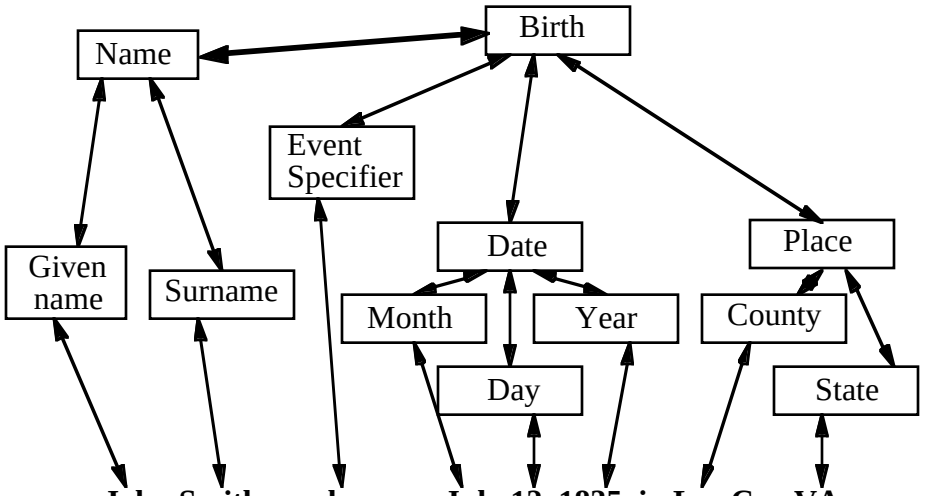
Suppose that a record contained the text:

```
John Smith was born on July 12, 1835, in Lee Co., VA.
```

Tags somewhat like the following could then be generated with the user's help:

```
<NAME ID=Name1, REF-FROM Assert1>
      <GIVEN-NAME>John</GIVEN-NAME>
      <SURNAME>Smith</SURNAME>
</NAME>
<ASSERTION ID=Assert1, Name1 HAS Birth1>was</ASSERTION>
<BIRTH ID=Birth1, REF-FROM Assert1>born
   <BIRTH-DATE>on
      <DATE>
         <MONTH>July</MONTH>
         <DATE-DAY>12</DATE-DAY>,
         <YEAR>1835</YEAR>
      </DATE>
   </BIRTH-DATE>
   <BIRTH-PLACE>in
      <COUNTY>
         <COUNTY-NAME>Lee</COUNTY-NAME>Co.
      </COUNTY>,
      <STATE abbr-for "Virginia" >VA</STATE>
   </BIRTH-PLACE>
</BIRTH>
```

The diagram below shows the same structure as the tags, but in a graphical format.



Note that the original text is still there, but each piece is tagged to indicate what it means. Tags are often hierarchical, so that, for example, a date is broken down into its month, day and

year. Tags can reference each other, e.g., in this example an "assertion" references the name ("Name1") and the birth event ("Birth1"), which in turn keep track of the fact that they are referenced from the assertion ("Assert1"). This would be an example of a low-level conclusion that would likely not be much disputed (i.e., the assertion is not saying anything about whether John Smith was really born then and there, but rather it is simply saying that the birth event within this statement is structurally attributed to John Smith, which is clear from the sentence structure).

Many tags can be generated automatically. For example, if the user highlights a date, the software could usually recognize the day, month and year and take care of tagging those automatically. The user should almost never need to see the tags explicitly.

The tags could be done in XML, or the same information could be stored (perhaps more efficiently) in a relational database of some kind.

Link to other data. Once the structure of the data in the record is extracted, the elements in the record can be linked to elements in other records or databases.

For example, places can be linked to (and from) a *place authority*, which is a database that contains further information on the location, history, etc., of each place [3], so that information beyond that actually appearing in the document does not need to be repeated in many places, e.g., "Lee" can be linked (bidirectionally) to a master entry for Lee County, Virginia, USA, which tells of the location of the county over the course of its history, the towns within it, links to all other records that reference it, etc.

As another example, given names can be linked to a name database that has information about how names and nicknames are related (Mary/Polly, William/Bill/Billy/Will/Willie/Wm., etc.), taking into account the time period and location.

Assertions can be made regarding other references to someone with the same (or similar) name that might be referring to the same person. Some of these assertions might link individuals in records or assertions to people in the user's own personal database, which hopefully will already be made up of individuals who link to other real records, all of which may already be shared with the master library (except perhaps for copyrighted or private information). Again, any such assertions point to the records they connect, and are also *pointed to* by those records.

As mentioned above, software algorithms can do much of the simple tagging automatically. The software can also suggest likely links, which the user can then verify. As the algorithms become more sophisticated, they will be able to do more and more of the work automatically, once they become at least as accurate as humans at each task. Human-generated (or supervised) data can be used to train and test learning algorithms and statistical models to allow much faster extraction and linking of data.

Every tag, assertion and link carries with it information about where the information came from and who entered it. Thus it should be possible to follow a link to see the user name or to see what particular version of an automated algorithm was used to enter certain data or to make assertions. If some algorithms (or users!) turn out to be unreliable (or even malicious), their contributions can be treated as such (or removed, if needed, leaving evidence and conclusions entered by others intact).

At every step in the above process, if the information exists, the user skips the step; if not, they can do the work and add the result to the master repository so that it is done. After the above steps have been done, the next person who comes across this record should be able to traverse the structure and links to find all related information that has already been found for this record. If they have additional information to add or to link in at any of these steps, they can do so.

5. Advantages

There are several advantages to using bidirectional links and evidence-based genealogical databases. First of all, every piece of data in a well-done genealogical database would have links showing exactly where it came from. (In fact, you could theoretically remove every name, date

and place in your database and it should be possible to reconstruct them by following the links down to the actual text from which they were derived.) This allows much better verification and easier resolution of conflicts.

In addition, every piece of data in a record points to everything that references it, allowing you to see all of the work that has been done in regards to that information. This allows a genealogist to see where else a person was referenced, who they are related to, whether their temple ordinances have been performed, and what additional work remains to be done for them. This also prevents individuals from "falling through the cracks" and never getting added to the global database. Conversely, it prevents references to people from being linked incorrectly to multiple real people. (For example, a marriage record for John Smith might otherwise be linked to two different nearby births for two different John Smiths.)

Because of the links attached to source records, genealogists can avoid duplicating work already done by others, which allows them to actually contribute something permanent and worthwhile when they further the work. This allows collections of records (and eventually all except perhaps the current, most recently generated ones) to be completely assimilated into a unified genealogical database.

Finally, as more and more of the process is automated and duplication of effort is avoided, each genealogist may be able to increase their efficiency by orders of magnitude.

6. Further Work

While the potential benefits of this approach are exciting, there remains much work to do in order to make this approach practical for general users.

At a high level, additional thought and experimentation are needed to decide upon a general implementation approach. For example, the GENTECH Genealogical Data Model, XML and/or relational databases could be considered for implementing tags, assertions and links. Some work is currently being done to develop an XML implementation of the GENTECH Genealogical Data Model [2].

The "master library" could be designed to make use of a central repository or a distributed database (with redundancy for the sake of data permanence). Work needs to be done to decide how to uniquely identify records, how to easily locate them (especially to avoid duplication), and how users can add new records if they do not exist in the repository already. Storage space, servers and internet bandwidth as well as other resources would be necessary, especially for a centralized repository.

Specific attributes and legal values for each attribute would need to be specified for places, dates, names, etc. Name authorities, place authorities, and other such systems are needed in order to incorporate domain knowledge about each of these kinds of attributes. Similarly, there needs to be a standard way to make each kind of assertion in order to make assertions consistent and to avoid ambiguity and confusion.

This method would be much more practical if many of the original documents are scanned and the images are made available over the internet, in order to allow widespread assistance on transcribing, tagging, and linking of the information. Software also needs to be developed with user interfaces that make it easy for people to learn and use the method.

There needs to be a way to allow a smooth transition from current approaches to the evidence-based approach. There is also a need to figure out how to use all of the existing genealogical databases to help rather than slow down this effort.

Algorithms are needed for parsing more and more kinds of data (e.g., dates, census forms, computer-generated family histories, etc.) to allow users to spend their time on tasks that really require human guidance. OCR, handwriting recognition, text parsing, automated searches for likely links, etc., can all speed up the process. As more extracted data becomes available, it will become easier to develop and test such algorithms, which in turn will help accelerate the extraction of further data.

Privacy and copyright concerns need to be addressed. For example, a link to user's database could indicate that work has been done, but the link may not be accessible without special permission from the user. Perhaps private data could be archived with a trusted repository for

preservation until such a time as it would be safe to release it (e.g., 110 years after the birth of the individuals involved).

Finally, genealogists need to continue to preserve, gather and scan existing records, and generate new records as genealogy continues to "happen" around us.

7. Conclusions

While there are many people working on genealogy at present, much of their efforts are spent doing work that has either already been done or will likely be repeated at some point. By using evidence-based documentation and bidirectional source links, genealogists can avoid duplicating efforts, and can instead spend their time doing new genealogical work that can be used by everyone.

There are many details to work out before these ideas are ready for widespread use, but if done properly, such approaches should help people to do genealogical research tasks "once" and "for all."

References

[1] 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*. May 29, 2000. http://www.gentech.org/gdm/

[2] Fugal, Hans, "An XML Implementation of the Genealogical Data Model", November 11, 2001. http://students.cs.byu.edu/~fugalh/gentech01/.

[3] Feist, Rob, "Place Database," *Workshop on Technology for Family History and Genealogical Research (FHT'01)*, Brigham Young University, March 29, 2001. http://www.fht.byu.edu/workshop01/