

Refocusing Family History Software And Capturing Research Intent

Chris Chapman

Abstract

The coming forth of distributed computing and modern genealogical research methods, such as the Genealogical Proof Standard (GPS), have caused us to rethink some basic assumptions about the design of traditional family history software. Most current family history software focuses on saving and retrieving the *current state* of the genealogical research. The research process is something that the researcher must keep track of separately, in research logs and other notes. However, moving the research process into center focus in family history software will result in better collaboration among researchers and more efficient use of research time. What needs to be modeled in family history software is not merely the vital facts, but the research process and intent.

This paper is not intended to be a rant about the current state of family history software, but to outline how family history *research* software could help the researcher by acting as a natural part of the research process. Many of the problems of database-driven family history software exist because of the focus on data management rather than process management. Software that is driven by the research domain instead of data management will simplify the tasks of the researcher and move the emphasis from bookkeeping to quality genealogical progress.

Contents

1	Family History Reporting	2
1.1	Definitions	2
1.2	The Beginnings Of Data Reporting	2
1.3	Data Collaboration	3
1.4	Limitations Of Data-Reporting	3
2	Moving Beyond Data	4
2.1	Research Reporting	4
2.2	Domain-Driven Family History Research Software	5
2.3	Research Collaboration	8

1 Family History Reporting

1.1 Definitions

There are a few terms that I use in a special way in this paper. I think it will inflict the least amount of pain if I go over the definitions at the beginning. I use the term *reporting* to mean the act of sharing some aspect of a genealogical research process with others. This definition is similar to enterprise reporting, in which reporting describes some aspect of a business process. In this paper we will cover two types of reporting:

Data Reporting A type of reporting in which the *end result* of the research process is shared. This can include source citations that support the conclusion. This is what most of us are used to seeing at this point in time.

Research Reporting A type of reporting in which the *entire* research process is shared. This will be covered in section 2.1.

I am also using the terms *family history* and *genealogy* synonymously. A *genealogist* is to genealogy as a *family historian* is to family history. I am using the term *family history software* to mean any software that can be used to maintain a family tree in a way that is semi-disconnected from the FamilySearch API (it should be able to at least store metadata about API data objects). Parts of this paper may be of use to products that rely completely upon the API for data, but I am not specifically targeting those in this paper.

1.2 Family Bibles And The Beginnings Of Data Reporting

From the beginning, people have tried to keep track of their families. Names and dates of descendants and ancestors were written down periodically so that a history could be passed down from generation to generation. At some point, this information was written down in books—in family bibles. As the printing press was developed, printed forms were introduced with specific fields for the needed information. Governments and churches instituted forms, or certificates, for the birth and death of their constituents. Some of these collections have been published and sent to libraries or other archives for others to read and research. Most likely, each of us has our own personal stash of these forms, that prove our birth and other vital information.

Family historians collected these forms for their deceased ancestors to prove kinship and other vital information about their ancestors. Once a historian has come to a conclusion about something, the conclusion was normally reported to others. Most reporting up to this point has been data reporting (see section 1.1). Over time, more printed forms were developed, such as family group records and pedigree charts, giving us a common way of sharing research conclusions. These forms contain fields that list all the vital information for an individual or family. At first, these forms had to be painstakingly filled out by hand. When the typewriter came around, it helped, but still it was a slow process, as each form had to be filled out one at a time.

Electronic Data Reporting

In the 1980's, with a more widespread use of relational databases and the arrival of personal computing, it became possible to store the current state of individual research conclusions on a computer! This was a huge step forward in productivity, as you only had to enter each person's

vital information into the computer *once*, and you could print out as many forms as you wanted with little effort! This made it much easier to share genealogical data. Programs were developed to automatically sort and shuffle data into the appropriate database table. Historians could print out and share much more than they were able to before.

Another step in productivity came with the idea of an *electronic* data report. With a printed data report, the recipient of the report had to re-enter the data into their personal computer. Electronic data reports made it possible to skip the re-entering step. The computer could understand the electronic report and update the database automatically! The Genealogical Data Communications Specification (GEDCOM) was developed to encourage interoperability between various genealogy programs. GEDCOM essentially describes a common electronic report format that any program can use to import and export data.

Older paper forms and artifacts began to be digitized and saved on computers as images by historians. There are many artifacts which are shared that way today which would not be available to the public otherwise. Metadata is extracted from digital images and indexed for easy querying in research. There are vast repositories of this data available today. The Internet has made it possible to share this data easily. The Internet has led us to the next big step forward: Data Collaboration.

1.3 The New FamilySearch and Data Collaboration

Data Collaboration uses data reporting (see section 1.1) to collaborate on a mutual family tree. It can be as simple as exchanging pedigree charts with a fellow researcher or as complicated as the New FamilySearch data services. The purpose of data collaboration is to spread the load and to get more done faster. Since we are dealing with the current research state in data reporting, it is usually easiest to assign branches (from the tree) to the researchers. This provides some protection from duplicating research work.

The New FamilySearch data services were created as a way to share genealogical data more easily. The purpose of these new services is to act as a sophisticated data collaboration environment for all the records of potentially all the researchers in the world. These web services allow third-party tools to access and commit data continuously, which creates a continuous merging environment. You can be notified of changes that other researchers have submitted and can update your records instantly. This is a great improvement over GEDCOM as an integration method. Integrating with the New FamilySearch data services is a much more robust and safe solution than using GEDCOM for integration. The New FamilySearch data services provide an excellent foundation for the sharing and merging of genealogical data. This is the foundation upon which the next generation of family history data collaboration will be built.

1.4 Limitations Of Data-Reporting

While data reporting and collaboration are wonderful, they do not provide all the answers. Currently, a small part of the genealogical research process is handled by genealogical software, and a majority of the process is handled outside the software. Most, if not all, of current family history software products only retain the *current state* of the research. Researchers have been on their own to learn and practice good research processes. Most current family history software is really a database at heart.¹ Most have a template-like form into which you enter data. These data-driven

1. Elizabeth Shown Mills, "Research Reports," in *Professional Genealogy*, ed. Elizabeth Shown Mills (Genealogical Publishing Company, 2001), 355–357.

applications are an extension of the earlier paper-based reporting forms. Those inexperienced in good research processes tend to spend a lot of time at the computer entering inconclusive information.² To make things worse, many inexperienced family historians have used these printed reports, which are designed for the reporting of conclusions, for their research. This has led to a muddling of the distinction between genealogical research and genealogical reporting.

Disciplined researchers learn good methodologies like the Genealogical Proof Standard (GPS).³ These researchers have made sure not to enter anything into their software unless it complies with the GPS. All of their preliminary research has to be kept somewhere else. It has been said that, “There is . . . a large gap between the way someone researches information and the way they enter it into their genealogical data management system.”⁴ This shows that there is an inherent mismatch between how good research is done and how current family history software works. It is evident that current genealogical software is not meeting the needs of the genealogical research domain.

Why is there this gap between how family history research should be done and how family history software works? Wouldn't it be better if the software modeled the research process? If it did there would be no mismatch between the research and the software. External factors may have influenced this, such as the lack of effective mobile computing since, until the last few years, it has been easier to work with paper logs when out in the field than with computers. Smart and small mobile devices are beginning to change that. Long-distance collaboration is becoming more and more common because of the ubiquity of internet access.

Data reporting is an inhibitor to collaboration in some cases. Some highly-trained genealogists are reluctant to share their research because there is no method currently to share data in a GPS context—using Evidence, Proofs, and Levels of Confidence. If family history software was built on the GPS, new researchers could learn good research processes automatically, since they usually learn whatever their family history software teaches them.⁵ Highly-trained genealogists would be able to share their data more freely with others, knowing that it would be understood and valued in the proper context.

2 Moving Beyond Data

2.1 Research Reporting

Research Reporting, as was explained in section 1.1, involves sharing the whole research process, not just the current state of the research. What is the research process? The research process involves everything from locating a Source to writing a Proof that describes what Evidence you based a particular decision on. It is the process of following the GPS for each Assertion that you make.

A research report should enable others to see what Sources you checked, which pieces of Information from those Sources you used in your Analysis, and a living, working Proof. The Proof

2. This was me when I was fourteen. After entering about 4,000 individuals into PAF, I realized I had no idea where I got most of my information. I had not entered any sources.

3. Elizabeth Shown Mills, *Evidence Explained: Citing History Sources from Artifacts to Cyberspace* (Baltimore, Maryland: Genealogical Publishing Company, 2007).

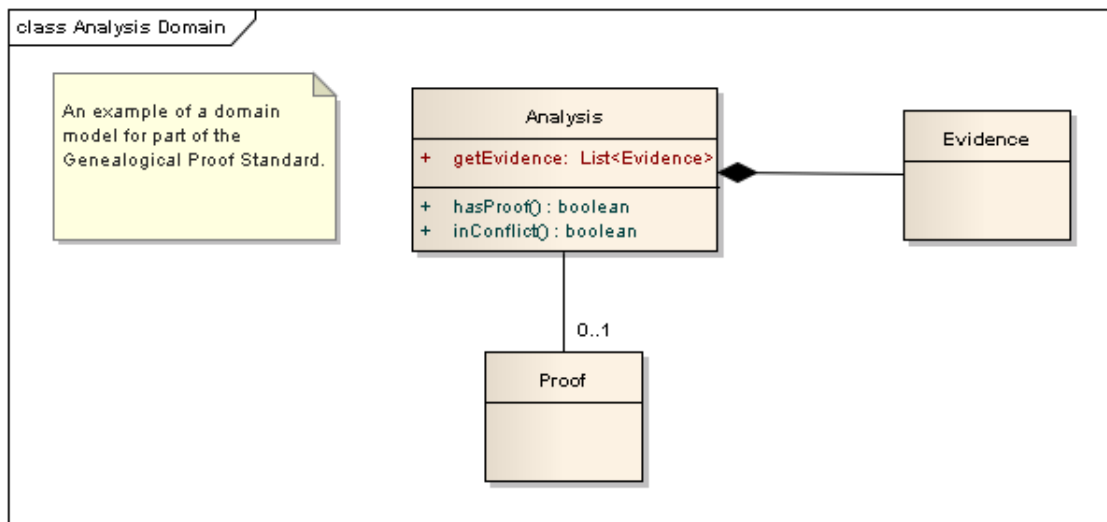
4. John Finlay, Christopher Stolworthy, and Daniel Parker, “Collaborative Research Assistant,” in *Family History Technology Workshop* (Provo, UT: Brigham Young University, 2007), http://fht.byu.edu/prev_workshops/workshop07/papers/1/collaborative-research-assistant.pdf (accessed March 8, 2010).

5. Mark Tucker, “10 Things Genealogy Software Should Do,” in *Family History Technology Workshop* (Provo, UT: Brigham Young University, 2008), http://fht.byu.edu/prev_workshops/workshop08/papers/1/1-3.pdf (accessed March 8, 2010).

should detail any Conflicts and how they were resolved. Just like in the GPS, there is no such thing as a final conclusion. Others should be able to build on the research that someone else has started and further refine our knowledge of the past, without duplicating research that was done before.

2.2 Domain-Driven Family History Research Software

While this paper is not intended to discuss implementation, I do want to give an example of how this could be done in a family history application. One type of design methodology that I deal with often in enterprise applications is domain-driven design (DDD).⁶ DDD focuses on modeling “real world” things and relationships in code. What is the core domain that we need to model in family history software? It is the GPS. The GPS is the process that we follow to establish a correct family history. This process is essentially followed for each Claim that we are trying to establish. Once the correct relationships of the GPS are modeled in the code, the complexity of the genealogical research process can be controlled. As an example, let us take one of the more simple relationships that we have in the GPS. This is not intended to be a full working model, but only an example of one way this could be modeled.



In this example, I have created an Analysis class, which represents the result of the third step in the GPS, where researchers “analyze and correlate the collected information to assess its quality as evidence.”⁷ The Analysis class contains a collection of Evidence objects, which in turn would know about the Source and Information extracted from the Source. The Analysis could also contain a Proof when Conflicts in the Evidence collection were resolved by the researcher. This is a very simple example, but hopefully can give an idea of how powerful modeling the research process could be.

The genealogical research process is very complex. If it were not, why would we be going to conferences and reading large books like *Evidence Explained*?⁸ A domain model centered around

6. Eric Evans, *Domain Driven Design* (Addison-Wesley, 2004), ISBN: 0-321-12521-5.

7. *The BCG Genealogical Standards Manual* (Provo, UT: Ancestry Publishing, 2000), ISBN: 0-916489-92-2.

8. Mills, *Evidence Explained*.

the GPS would be a very powerful thing. It would lead the researcher through accepted research patterns and practices. “Complexity in the heart of software has to be tackled head on.”⁹ The best way to do this is not with a research assistant add-on to data-oriented family history software, but with new software that will use the GPS at its heart.

Capturing Research Intent

Keeping track of data changes, or data provenance, is not enough. Data provenance would allow us to look back and see what changes were made, but would not really give us an understanding of *why* the changes were made. Knowing the *why* about a change helps immensely in collaborative family history research. Having the user enter a textual description when they are changing something is usually not enough. If you are familiar with commit logs on common source control management systems you will know what I mean. If you are not familiar, let me just say that developers are not known for being very explicit in what they changed and why when they commit changes to fifty different source files! The system should not be dependent on the researcher to enter the reason for a change. The family history research application itself must be aware of the context and intent of the user, and know what to do automatically to capture intent.

In some newer collaborative family history software I have seen that the editing process is more granular. To edit an ancestor’s death record, for example, you click the “Edit” button next to their current death information, and you enter only the death information in a smaller form. This is a step in the right direction, but this is still a form over a database. It is not capturing intent.

We need software that understands the research process. When you enter in or change information about a person, there is usually something that prompted it. You may have found a new source, or realized that you had mistyped information the first time. There is always a reason. This reason is the *why*. In collaborative family history software, it is essential to capture this reason, as it helps others know what you were doing at the time you did it and why. It helps you when you come back to it in a year to know why you made that change. We need to get away from fill-in-the-blank forms and capture research intent. By following the GPS, you will find that a template form where people can enter raw information about a person’s life is not really needed. Here is the great benefit that will come as a result:

Once we can capture research intent, we can direct research.

You may want to pause and think about that. What do we direct researchers in? Researchers are directed through the GPS. Directed research makes it possible for a busy person to accomplish something, even if they only have twenty minutes per week to spend on it.¹⁰ Since every record in the system is built upon the GPS, the research state of all the research is constantly known to the system. The system could give you choices on what needed to be worked on next, according to some predefined rules.

9. Evans, *Domain Driven Design*.

10. Charles D. Knutson and Jonathan Krein, “The 20-Minute Genealogist: A Context-Preservation Metaphor for Assisted Family History Research,” in *Family History Technology Workshop* (Provo, UT: Brigham Young University, 2009), http://fht.byu.edu/prev_workshops/workshop09/papers/2-Lunch-Knutson.pdf (accessed March 8, 2010).

Source-Driven Research

The effect of designing a system based on the GPS is that researchers think about the sources first. If there has to be a reason to change something before it is changed, focus moves from the template-like form to the source. In most current software, you fill in the form first, and then you enter in the sources. Isn't this backwards from how it happens in the field? As every good genealogist knows, you find sources first, then you make your conclusions.¹¹ We gather information, then analyze it to see whether it is relevant. We resolve conflicts that we find until we can make a strong conclusion, with a written proof. Family history software should do the same thing. The system should not get in the way of our thinking, but it should bring pertinent evidence and information to our attention so we can make sound conclusions.

Rethinking User Interfaces

To capture research intent, many changes need to be made to the traditional user interface. Here are some problems that we currently have when entering data:

The data entry process can become laborious. As an example, if a researcher finds a census record which enumerates 6 individuals, they now have to navigate to each of those 6 individuals and enter in multiple fact records for each person. At the same time, they must ensure that source citations are properly and consistently entered. Genealogy researchers need a way to simplify their translation of research results into genealogical data. They also need a way to simplify the entry of that data while including proper source citations.¹²

We can do better than this. According to the GPS, the first thing that we need to do is to “conduct a reasonably exhaustive search in reliable sources.” This should be the first thing that we do in our research system as well. We need to get Sources. The easiest way to import a Source into our research system would be to scan or photograph it. From the resulting image, metadata could be extracted.¹³ Filing cabinets and libraries would still be needed to store the record that was imaged for source provenance. Up to this point, we have only entered a Source into our system. The Source is not related to anything yet. Once the metadata was extracted and the Source was Cited, we could relate the Information from the Source to some Assertion on some person or persons. It is possible that the Source described someone whom we did not have in our system yet. The name on the image could be highlighted and a context menu could allow you to create a new person from the Source. The system would be aware of which Source you were viewing, so you would not need to add any additional source records anywhere else. All data entry starts from source image metadata from fields on imaged Sources. This is the same way you would discover information following standard research processes.

Let's look at a more complicated scenario. Let's say a researcher has found four different Sources for his fourth great-grandmother's birth date. The Sources are in the system and have been related to the birth of the correct person. The system notifies the researcher that two of the four Sources show a different birth date. Now the researcher has an opportunity to resolve Conflicts, just like in manual research processes. If the researcher decides to resolve the Conflict now, he is taken to

11. *The BCG Genealogical Standards Manual*.

12. Finlay, Stolworthy, and Parker, “Collaborative Research Assistant.”

13. Tucker, “10 Things Genealogy Software Should Do.”

a screen where the four Sources are listed and weighed. The researcher can Analyze the known Sources or add additional Sources. After making a decision, the researcher can write a Proof that explains the Conflicts and the decision that was made.

So, in our modern user interface, Sources should be entered first. Then they can be related to people. An Assertion about a person is the end result of all the Evidence collected and a Proof statement. The Proof statement should explain any Conflicts or problems encountered in the decision and include the reason why a particular piece of Evidence was accepted or rejected. The whole user interface is built around research intent.

2.3 Research Collaboration

Research Collaboration would use research reporting (see section 1.1) to collaborate on a mutual family tree. It can be done manually, using research logs, but is most easily done with a computer application that understands research intent. It could be done in a multi-user collaborative system with great potential as well as being synchronized through an external service, such as FamilySearch. The purpose of research collaboration is to spread the load and to get more done faster. It is much more efficient than data collaboration (see section 1.3) because the research would not need to be duplicated to comply with the GPS. The entire research context would be available.

Research collaboration would generally work best in small family groups. It allows a much more flexible breakdown of duties than with data collaboration. It would be possible to collaboratively research the same individual without much fear of duplicating work.

Conclusion

If family history software models the research process, it will become a much more powerful tool for the genealogist. Novice genealogists can learn correct research principles and processes from their family history software. If the research intent can be captured by software, the research can be directed and saved. A template-like form used to enter information about a person does not capture research intent. Once software can capture research intent, other researchers will be able to understand the Evidence that led to certain conclusions. When new Evidence arises for a particular Assertion, the system will notify the researcher to decide whether the new Evidence changes any previous conclusions. If it does, the software will know that the new Evidence resulted in a change in a conclusion. Because the user interface is focused around research intent, many of the problems that we have had with traditional genealogical data entry will be eliminated. Research progress will skyrocket as researchers shift their focus from bookkeeping to creating a reconstructed family history that is as close to the truth as possible.¹⁴

References

Evans, Eric. *Domain Driven Design*. Addison-Wesley, 2004. ISBN: 0-321-12521-5.

Finlay, John, Christopher Stolworthy, and Daniel Parker. "Collaborative Research Assistant." In *Family History Technology Workshop*. Provo, UT: Brigham Young University, 2007. http://fht.byu.edu/prev_workshops/workshop07/papers/1/collaborative-research-assistant.pdf (accessed March 8, 2010).

14. *The BCG Genealogical Standards Manual*.

- Knutson, Charles D., and Jonathan Krein. "The 20-Minute Genealogist: A Context-Preservation Metaphor for Assisted Family History Research." In *Family History Technology Workshop*. Provo, UT: Brigham Young University, 2009. http://fht.byu.edu/prev_workshops/workshop09/papers/2-Lunch-Knutson.pdf (accessed March 8, 2010).
- Mills, Elizabeth Shown. *Evidence Explained: Citing History Sources from Artifacts to Cyberspace*. Baltimore, Maryland: Genealogical Publishing Company, 2007.
- . "Research Reports." In *Professional Genealogy*, edited by Elizabeth Shown Mills, 355–357. Genealogical Publishing Company, 2001.
- The BCG Genealogical Standards Manual*. Provo, UT: Ancestry Publishing, 2000. ISBN: 0-916489-92-2.
- Tucker, Mark. "10 Things Genealogy Software Should Do." In *Family History Technology Workshop*. Provo, UT: Brigham Young University, 2008. http://fht.byu.edu/prev_workshops/workshop08/papers/1/1-3.pdf (accessed March 8, 2010).