Efficient Genealogy through Personal Extraction and Automated Verification

D. Randall Wilson
Fonix Corporation
Draper, Utah
WilsonR@fonix.com

1. Introduction

Traditional genealogical research often involves (1) searching for documents containing information about relatives, and then (2) entering the extracted genealogical information from the documents into a genealogical database. Often the overwhelming majority of the time is spent on the first step, because documents must be sought at libraries, on microfilm, or even at distant locations. *Extraction* efforts involve entering all of the genealogical information about all of the people referenced in a source into a database. If extraction is done *before* the searching phase, then searches can be completed in seconds instead of hours.

Volunteers are more likely to be interested in helping with such efforts if they are able to choose assignments of personal interest to them, and also if they know that there is a "critical mass" of volunteers that are all pitching in to extract a significant portion of the available relevant records.

Technology has been of great benefit in helping researchers to organize and share their research and conclusions with others around the world. However, it is important to make sure that both traditional research and extraction efforts all contribute to the global effort of building a worldwide genealogical database instead of simply multiplying the number of (often undocumented) copies—and perhaps erroneous variations—of information taken from original sources. Furthermore, there remains a need to be able to automatically verify work done by others so that the work does not have to be duplicated.

This paper discusses what makes extraction approaches efficient, and how personal extraction provides additional motivation for people to participate. It then shares ideas on the harder issues of how to use the extracted records as the basic building blocks of a new kind of evidence-based genealogical database. Finally, it discusses approaches to verification in order to provide some hope of being able to trust the work of others and build a truly collaborative global genealogical database.

2. The Efficiency of Extraction

Searching for a *particular* ancestor (or other relative) can take a long time—hours, days and sometimes more—and can be very expensive. There may be little evidence to guide the researcher to know what books or records to look through, requiring much fruitless hunting. Often books and other records lack an index, making it extremely time-consuming to look through the book to see if there is a mention of their relative. Microfilms can be cumbersome to scroll through, and yet the alternative—going to a distant county or country to look at the original documents—would be prohibitively more difficult in most cases. Most documents have not yet been digitized to the point of allowing efficient computerized searches for individuals referenced within them.

A researcher will typically pass over hundreds or thousands of names in the search for one of their own relatives. Genealogical research could be made much more efficient if people worked together to extract all of the relevant information from entire documents (or collections of documents) at once instead of spending so much time searching for specific pieces of information. As the volume of extracted information grows from everyone's efforts, locating documents referencing one's own ancestors would become much faster, and using the information would be easier because the structure of the information would already have been extracted.

In this paper the word *document* is used to refer to a source containing one or more pages (though it could sometimes apply to electronic information such as a GEDCOM file). The word *record* is used to describe a data structure containing information taken from a reference to a person in a document. A document can therefore have as many "records" as it has references to people. Note that a record does not represent a *person*, but a particular *reference to a person* as it appears in a document, similar to Harten's "identity" (Harten, 2002), and Gentech's "persona" (Gentech, 2002).

Extraction is more efficient than the more typical "search first" method of genealogy for several reasons:

- Once extractors have entered one *record* (i.e., a name and the information about it), the next entry is immediately after it, so there is almost no search time involved.
- Once extractors learn how to properly enter the information for one record, they already know how to enter the information for the next one, so training time is decreased.
- Having just entered one record, the next one goes more quickly because the process is fresh in their mind (like an assembly line), and they know where each piece of information goes without having to think too long about it.
- Once they have specified the citation once for a document, the documentation of each individual record is almost completely automatic.

Searching will eventually have to be done, of course, but if the extraction is done *first*, then the searching can take a few seconds per record instead of minutes or hours. Because the information in the extracted records are highly structured, more accurate searches are possible with less "false matches" to wade through than when searching through a simple index or transcription.

In addition to the efficiency, there are several other advantages:

- Since extractors know they will be entering many of the same kind of record, they will be more likely to take the time to figure out how to enter it correctly.
- Since one person will be entering many records, they will tend to be entered more consistently than if a different person entered each one.
- The work is straightforward, so it is not confusing to users, and the raw extracted data can easily be stored in a structured form that will allow for automated conversion to whatever format is needed in the future.
- Since everyone on a page has their information extracted, nobody "falls through the cracks."
- By simply keeping track of which records have been extracted, duplication of effort can be largely avoided (except limited duplication for the purpose of verification).

While advances in such technology as OCR, handwriting recognition, natural language processing, parsing of semi-structured data, and other such areas may aid in the work of extraction, humans are still superior at being able to look at almost any kind of document and determine what genealogical data is contained in it, and how it relates to other information in the document.

3. Personal Extraction

Some extraction work is already being done, of course. However, people are more likely to volunteer to help with such efforts if they are allowed to choose documents of interest to them to work on, such as those that involve an area where their relatives are from. For example, if a researcher has ancestors from Lee County, Virginia, then they might be willing to spend some time doing extraction work on census records from that county. Whether or not they extract information on their known ancestor, they would know that the people whose information they are extracting could be neighbors, friends or even relatives of their own ancestors, the connection to which might be facilitated in part by their own extraction efforts.

Such extraction efforts would be most effective if done on a massive scale, with many volunteers helping with the work. That way each person would be able to spend about as much time as they would finding a few relatives, but instead make 10 or 100 times that many individuals available for themselves or others to find much more quickly. When people know there are lots of other people doing extraction, too, it will help them feel like they can both contribute to and benefit from the work being done.

Internet extraction tools. For such a *personal extraction* effort to be practical, it would be necessary to develop tools to allow such work to be done over the internet. Digitized images of documents would need to be made available. Initially enough images would need to be made available to keep volunteers busy. Then additional scanning could be done to stay ahead of the volunteers. Users could either work directly on-line or they could download images, work off-line and upload the extracted information.

Initially the tools might have templates built-in for certain kinds of documents (birth records, census records, etc.), and could make use of algorithms to automate certain aspects of the work, such as identifying the rows and columns in the document (Nielson & Barrett, 2001), and could automatically tag the information taken from each field. Less structured documents could also be supported using a more flexible set of data primitives.

Choosing projects. Users could choose from a list of projects currently available from the server, choosing the location and/or type of record they want to work on. They might even opt to start with a page they know has reference to their own relatives, and then go on from there. They could also be allowed to do extraction on documents they have in their possession already, and add the extracted records (and perhaps an image of the document) to the extraction database. Volunteers who are willing to simply help with whatever needs to be done could be assigned tasks based on priorities assigned by administrators. For example, census, birth, marriage and death registries might be given priority over land records.

The server could keep track of which pages of which sources have been assigned out, and would gather the extracted data back together. (It might even send out reminders to those who have had something "checked out" for a long time and/or let them know when their volunteered page(s) have been reassigned).

Traditional research. There are times, of course, when extraction is not the best approach to doing genealogy. In particular, researchers typically must do traditional research on living individuals and other close relatives by themselves because privacy concerns would prevent the open sharing of records referencing these people. However, even then they could build the same kind of records built by extraction efforts—in essence, "extract" the information about their own relatives—and add it to their own local database. Then they could wait to share it until it is safe, or else to make sure it is preserved they could submit it to a trusted repository that will leave it in a "time capsule" until it is safe to release. Alternatively, the repository could share just enough information with the public to allow people to contact them personally for more information.

The remainder of this paper deals with what to do once records are extracted from original sources, whether they come from batches of extraction efforts or simply from extracting the information from a single entry in a document after a more traditional search.

4. Fundamental Inefficiency Issues with Current Methods

Once these records are extracted, they can of course be used to speed up searches, and currently the data would typically be copied into personal genealogical databases. However, most current genealogical databases suffer from several fundamental problems. They "take a shortcut" and "deal with only tidy, final conclusions...sidestepping the harder job of storing, linking and sharing evidence, and of representing conflicting and changing opinions" (Harten, 2002). They typically do not link with much precision back to the original sources, do not always resolve which fields came from which source, and lack rigid citations, which make it very difficult for computers to determine if two pieces of information came from the same place. Furthermore, "a conclusion-only database offers no basis for future evidence evaluation: no researcher can build on the work published by another" (Harten, 2002) Because of these and other problems, current genealogical databases make collaboration (in the sense of working together to do the research once and for all) very difficult and make it highly likely that work will be duplicated and never really "finished" (Wilson, 2002).

The National Genealogical Society includes in its list of "Standards for Sound Genealogical Research" (NGS, 2002) the following statements:

> *Family history researchers consistently—*
> - *Record the source for each item of information they collect....*
> - *Seek original records, or reproduced images of them when there is reasonable assurance they have not been altered, as the basis for their research conclusions.*
> - *Use compilations, communications and published works, whether paper or electronic, primarily for their value as guides to locating the original records, or as contributions to the critical analysis of the evidence discussed in them.*

While these are sound principles for traditional genealogical research, they carry the disturbing implication that each researcher must verify the work done by anyone else by personally examining

the original sources. This again implies that "no researcher can build upon the work of others" (Harten, 2002) because no one can trust anyone else's research.

While GEDCOM files and published family histories do make research faster the second time around by guiding the researcher more quickly to the original records, they do *not* make the research faster the *third* time around: Any verification performed by one researcher typically has to be done again by every subsequent researcher. Certainly we do not want millions of people to all have to individually verify billions of records—that would be about a million times too much work. Rather, what is needed is a way to allow for automated verification of the work done by others.

5. Extraction records as basic building blocks

Part of the solution to the problems discussed in Section 4 may involve using extracted records as the fundamental building blocks of a new kind of evidence-based genealogical database. In such a database, each *record* represents not a person, but a *reference to a person* (also called an "identity" (Harten, 2002), or a "persona" (GENTECH, 2002)), as mentioned above. Each such record contains all of the information about the person referenced that is obvious *from the document itself*. Knowledge from other sources should not be used to modify the extracted record, but it should be able to stand on its own in such a way that researchers would agree that the extracted information is indeed what the record says (whether or not the record itself is actually correct).

Each record can also contain links to other extracted records from the same page. Multiple references to the same person in a document can be combined into a single reference when it is obvious that the same person is being talked about. When there is any doubt, then a separate record should be created and linked to the other one with the justification given for believing they are the same. For example, consider a paragraph from a printed family history that reads:

> *James Gray (b. 1800, VA) m. Louisa Parish (d. bef. 1850) and was the father of Carlo B. He died between 1860 and 1870.*

An illustration of what kinds of information an extractor might extract from this excerpt is given below in Table 1.

Record 1 Unique ID=[SourceID][Page#][Record 1] Name=James Gray=>James /Gray/ {common first name; common surname; common order of name} Birth date=1800 Birth place=VA =>Virginia {Common abbreviation} Married to: [Record 2] Father of: [Record 3] Sex: =>male {use of "father"; common male given name "James"} Extracted by: AnitaNOLA, 2 March 2003 Verified by: GrannySmith, 2 April 2003 Verified by: JimmyBoy98, 3 April 2003	Record 2 Unique ID=[SourceID][Page#][Record 2] Name=Louisa Parish Married to: [Record 1] Death Date=Before 1850 Extracted by: AnitaNOLA, 2 March 2003 Verified by: GrannySmith, 2 April 2003 Verified by: JimmyBoy98, 3 April 2003 Record 3 Unique ID=[SourceID][Page#][Record 3] Name=Carlo B. => Carlo B. Gray {because of father's surname} Extracted by: AnitaNOLA, 2 March 2003 Verified by: GrannySmith, 2 April 2003 Verified by: JimmyBoy98, 3 April 2003 Record 4 Unique ID=[SourceID][Page#][Record 4] Pronoun=He Died: Between 1860 and 1870 Same-as: =>[Record 1] {Subject of previous sentence; age 60-70 at death is common}. Extracted by: AnitaNOLA, 2 March 2003 Verified by: GrannySmith, 2 April 2003

Table 1. One possible representation of extracted records.

Note that each record has a unique ID by which it be referenced by the other records in the same document. A unique identifier is necessary for each record extracted from each source. This allows genealogical databases to precisely cite the source of their information, which in turn (a) allows computers to automatically tell if two pieces of information came from the same source; (b) allows databases to contain only references to the original records instead of having to always contain a copy of the information; and (c) allows researchers to more quickly access the image or transcription of the original source, if available, for the purpose of verification. It also allows the automatic generation of bidirectional source links (Wilson 2002) to help those looking at the extracted records to know what has been done with them.

In several cases the symbols "=>" are used to indicate an assertion that is fairly obvious from the record (e.g., that "VA" means "Virginia, U.S.A."). In each such case the justification for such an assumption is given in curly braces ("{}"). Less obvious assertions and links to other records are made in separate "assertion" records (not shown).

Note also that there is a slight ambiguity as to who the word "He" refers to in the second sentence. This warrants splitting "He" into its own record, which is then linked to Record 1 with a brief explanation of why the extractor thought these two records were the same. Later evidence might override the justification in the "Same-as" field (such as if a death record for "Carlo B. Gray" is found in 1865, and James Gray appears in the 1880 census).

Linking records together. Extracted records can serve as the basic building blocks of genealogical databases, but it is important to know how to link them together in order to identify which records likely refer to the same person. As suggested by Harten (2002), a "best view" could be constructed by linking (or "tying") several extracted records together into a single *person record*. The fields of a person record are not typed in, but come directly from the extracted information (perhaps with some amount of value normalization for presentability). When there are conflicts, information about the sources could allow automatic prioritization (e.g., earlier, more specific and more primary sources over later, less specific derived sources). Users could rearrange the default priority by again providing proper justification for doing so. The actual information from each of the original records could be copied to a user's personal database as a "locally cached" copy, but would not be subject to modification, since it would simply represent what was in the global database (though additional assertions about the record or even assertions regarding incorrect extraction could be made).

Not all data records need to come from extraction efforts. When users have further information about a person, they should create a new record and cite the source of the information (which could be an interview with someone, a document, book, photograph, etc.). They could check the global database to see if that source is already there, and add it if it is not. Then they would link the new record to another existing record, giving the justification for believing they refer to the same person.

This approach of linking original extracted records together allows users to trace any piece of information back to its original source. It also allows conflicting opinions to be explicitly stored instead of being replaced with "false simplicity."

Records can contain relationship links to other records (e.g., a child in a birth record would have a father and mother link to the parents' records, and they would each have a child link to him or her). Thus, when two records are linked together to indicate that they refer to the same person, then the combined "person record" contains all of the relationship links of any of its sub-records. When there are conflicts between such claimed relationships (e.g., links to two sets of parent records that are known to not reference the same people), then the records with higher preference can be used in generating the "best view" commonly displayed to users, as is done in choosing which records to use when displaying fields that may have conflicting possible values.

On the other hand, often when two records are linked, it will be common to examine the parents and other relatives to see if they, too, should be linked. Recursively following this process can potentially unite large databases of records that reference the same individuals (Wilson, 2001), and can begin to unite the work already done by so many people.

The decision to link two records together (e.g., to say that they represent the same person, or to say that there is a relationship between them) requires documented logic, or *justification*.

6. Justification

For any assertions or conclusions drawn from the extracted data, the justification for such decisions should be given. A *justification record* might include an asserted fact, links to one or more other records (or even fields within the record) to support the justification, a stated reason to believe what is being asserted, and an identification of who made the assertion (e.g., a user name or algorithm ID), and a date.

Justification is needed for transforming data as written to another form (e.g., "VA"->"Virginia, U.S.A."), for inferring information not explicitly present in the data, for choosing a field for the "best view" from one record over another, for choosing one relationship link over another conflicting one, and especially for linking records together (i.e., for claiming that they both apply to the same person). Some common reasons that might be used to justify linking two records together

include:

- Similar names (e.g., "James Gray", "Jim W. Gray").
- Similar places (e.g., Harlan Co., KY, and Lee Co., VA, which are adjacent counties).
- Consistent time periods (e.g., "Bought land in 1850" and "age 35 in 1860 census").
- Same relatives (e.g., "Wife named Polly, son Jed.", "Wife Polly, sons Jedidiah and Sam").
- Same occupation, religion, race, place of birth, handwriting, etc.

Also important is the inclusion of reasons that the records might *not* refer to the same person, such as mismatches in any of the above, or too little evidence (e.g., "Mr. Smith" with no further information). It is also possible for a record to be linked to two different records that conflict with each other, indicating that at least one of the two links is not right.

When extracted information is entered, the implied justification for all of the information given is that the information was written in the original document. The source should be identified as a *primary* source (i.e., written by an eyewitness of the event, preferably close to the time of its occurrence) or a *secondary* source (i.e., something other than an eyewitness account, such as a birth date recorded on a tombstone). *Derivative* sources can be specified, too, in which it is acknowledged that the information was taken from another known source in order to make clear that these are not two independent pieces of supporting evidence, but that one came from the other. When the other record has also been extracted, then it would be likely that these records would be linked to each other.

Often there will be multiple copies of a source. For example, there might be several copies of a book stored at different libraries, microfilmed copies of original records, etc. While each source has the potential of being different (e.g., a different edition of a book, a page missing from one of the copies, etc.), when it is likely that the entire source is repetitious, it does not make sense to repeat the extraction, but rather it makes sense to link the sources together such that work done on any of the copies becomes connected to all of the other copies.

7. Automating Verification

As mentioned above, there needs to be a way to automatically verify records and assertions that are contributed to a global database or into one's person database. Otherwise each researcher would need to personally verify everything in the database, which becomes impossible as the database grows large.

Computerized Verification. One approach to this problem is to try to make it possible for computers to do the verification automatically. This approach would require that the justification given for each conclusion be provided in a rigid enough format to allow for algorithms to deal with them effectively. For example, if a name is given as "James Gray", it would be common to use the order of the names in the region, as well as the commonality of the given and surnames to conclude that the given name is "James" and the surname is "Gray." Tags like "Common-Given-Name", "Common-Surname" and "Common-Name-Order" and many more could be defined to make such logic understandable by computer algorithms.

If this were encoded properly, then as more extraction is done, statistical data could be gathered about how often each kind of conclusion holds up, etc., which could aid in automatically choosing which of several conflicting hypotheses is most likely to be correct. Eventually this could help in automatically providing justification for automated linking and other such algorithms.

Signature-Based Verification. A simpler approach, especially in the short term, is to simply have other human researchers double-check the extraction and simple conclusions and add their "signature" (e.g., in the form of a verified username and timestamp) indicating that they have verified that the extracted record correctly represents the information in the original document. Once a couple of signatures have been added to a record, other researchers should be able to rely upon that part of the database.

Similarly, users could add their signature when they verify any other stages of the process or any other work that is done. For example, they could verify that a digital image accurately represents the original microfilm or record, that a transcription, if any, is valid; that they agree with the extraction; that they agree with an assertion or conclusion and the justification given, etc.

Where there are differing opinions and not enough evidence to know what is right, there may be

signatures that vote both for and against a given assertion, which would help researchers identify areas that have not reached a consensus.

If there is concern over whether the witnesses are trustworthy, then perhaps a rating system could be provided by which researchers gain more trust ("credentials") depending on how much work they've done, and whether there has been much disagreement with what they've done. Similarly, automated algorithms could be assigned a certain user ID and could be assigned a confidence level based upon their measured accuracy.

8. Conclusions

Doing extraction first and searching later is much faster in the long run than the other way around. Extraction work also provides complete coverage of the information in documents. Personal extraction is more compelling because users get to choose projects that are of interest to them. Tools that allow people to do extraction work over the internet would open the project to millions of volunteers.

A new kind of evidence-based genealogical database that uses records extracted from original sources could help overcome the "false simplicity" of genealogical databases by storing and dealing with conflicting records. Such a database would allow one to follow each conclusion back to the original sources it is based upon. By keeping track of users of a global extraction and linking effort, each step of the process—scanning, extracting, linking, and drawing conclusions—could be independently verified by enough people that it could be trusted by others, and areas of unsolved questions could be more clearly identified as such. Extracted records and justification for links and other decisions can be shared instead of just conclusions, thus providing a way to integrate everyone's work without having to make arbitrary decisions about conflicts.

There are a number of things that need to be done to put such ideas into practice, of course. There needs to be a system of providing unique identifiers to sources. There needs to be a global (central or distributed) system for delivering document images and collecting extracted records. Tools need to be built to make it straightforward for users to extract properly structured data and that attach the proper citations to each item.

The details of the structure of records, links, and justifications in evidence-based genealogical database also need to be worked out. This is not a trivial problem, but is a fundamental issue that must be solved before truly collaborative global genealogical efforts can succeed. With such a foundation, additional advances in technology and algorithms can do more than just help us to duplicate each other's efforts more quickly—they can help the work truly move forward at an accelerated pace.

References

GENTECH Lexicon Working Group: Anderson, Robert Charles, Paul Barkely, Robert Booth, Birdie Holsclaw, Robert Velke, John Vincent Wylie. *GENTECH Genealogical Data Model, Phase 1: A Comprehensive Data Model for Genealogical Research and Analysis*. May 29, 2002. http://www.gentech.org/gdm/

Harten, Bill, "System For Identity Linking and Research Collaboration," In *Proceedings of the 2nd Annual Family History and Technology Workshop (FHT 2002)*, pp. 61-63, 2002.

National Genealogical Society, "Standards For Sound Genealogical Research," 2002. http://www.ngsgenealogy.org/comstandsound.htm

Nielson, Heath, and William A. Barrett, "Automatic Zoning of Digitized Documents," In *Proceedings of the Workshop on Technology for Family History and Genealogical Research (FHT 2001)*, pp. 8-9, 2001.

Wilson, D. Randall, "Graph-based Remerging of Genealogical Databases," In *Proceedings of the Workshop on Technology for Family History and Genealogical Research (FHT 2001)*, pp. 38-39, 2001. http://axon.cs.byu.edu/~randy/gen/Remerge.html

Wilson, D. Randall, "Bidirectional Source Linking: Doing Genealogy 'Once' and 'For All'," In *Proceedings of the 2nd Annual Family History and Technology Workshop (FHT 2002)*, pp. 54-60, 2002. http://axon.cs.byu.edu/~randy/gen/Bilink.html