Family History Technology: A Survey of Ten Hard Problems

Dr. Doran Wilde
Brigham Young University, Dept. Electrical and Computer Engineering
dwilde@gmail.com

The ultimate and overarching aim of family history research is to build a common lineage-linked family tree for the whole human race based on available records. In the light of this extremely ambitious and seemingly impossible challenge, the hope has been that technology could help bring this goal within reach and even accelerate the rate at which progress is being made. To this end, the Family History Technology workshop was started in 2001 with the charter to promote research in the most difficult and unsolved problems that impede genealogical research. The workshop provides a forum for researchers to discuss their research and trade ideas on those hard, unsolved problems whose solutions could dramatically accelerate the work of family history.

In the past several years, we have seen an increase in the availability of source records and more and more of those records have been indexed. This has made a big improvement in how much family history can be done. However, many of the truly hard problems that impede genealogical work still persist and there are only a precious few who are working to solve those problems.

The objective of this paper is to restate the purpose of the Family History Technology Workshop and survey some of the hard, unsolved problems that need more attention and research. The hope is to refocus our workshop community on those key emerging technologies that could enable the important key breakthroughs needed to revolutionize the way we do genealogy and ultimately fulfill the promise of technology to accelerate family history work.

This paper will survey each of the following ten problems, give a brief synopsis of the problem and discuss the potential role of each in family history research.

1. Document Digitization and Preservation
2. Document Processing
3. Handwriting and Optical Character Recognition
4. Document Markup
5. Data Extraction
6. Family-Linked Data Models
7. Record Linking and Record Merging
8. Digital Experts & Assistants
9. Population Reconstitution
10. Intelligent Search, Visualization, and Human Computer Interface

These ten topics are tightly interrelated. They could potentially work together as parts of a futuristic genealogical document analysis system in which source documents are input into the system and reconstituted family-linked populations derived from those documents are output. Such a system seems far-fetched and nearly impossible. However, you can decompose its functionality down to the set of modules listed above. Figure 1 below illustrates how this system might be constructed from those modules and how they might interface with each other.

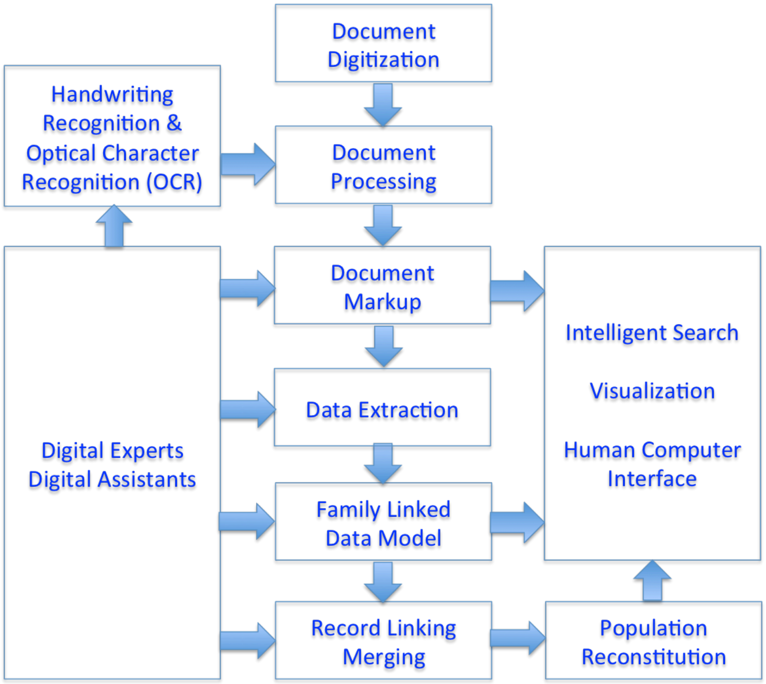

Figure 1. How the ten surveyed problems relate to each other in an overall record analysis system.

In the remainder of this paper, the functionality of these modules will be discussed, together with the technologies that need to be developed to make them a reality.

Document Digitization and Preservation

Digitization is the technology to make digital copies of records without losing information, and if possible, even enhancing the information contained in the record. Preservation concerns the technology needed to safely store records for a very long period of time (1000 years or more) guarding them against media degradation and failure, while at the same time making them available, accessible and searchable.

These technologies are employed by the entities that acquire, preserve and publish records such as the church, governments, libraries, and commercial genealogy publishers. This technology has seen a surge in just the last ten years. A decade ago, the primary means of preserving and publishing records was microfilm. The information content of the microfilm is not ideal. Color information was not recorded. Records that were difficult to read often became impossible to read on microfilm. Some microfilm images were captured out of focus, overexposed, underexposed or blurred. Now those films are becoming brittle, are deteriorating and are at risk of being lost. About five years ago, the church embarked on a project to digitize its collection of films, with the expectation that they can be preserved beyond the life of their original media.

Even records stored in digital formats are not safe. Digital formats evolve and change, and we can loose our ability to read the files. Digital media, like all physical media, is not problem free and files can get corrupted and be lost. Technology to address all of these problems needs to be put into place or we risk losing the records that have already been copied from their originals.

Document Processing

Document processing is the first step to trying to extract the information that is contained on a page of a digitized document. Often the spatial position of text within the document is as important as what the text says. Document processing is the technology needed to be able to spatially parse a page into meaningful parts to give context and meaning to what is written on that page. It includes spatial parsing of printed forms to locate lines, boxes, fields as well as the parsing of freeform text to identify blocks of text, lines of text, and spatial delineations between record entries. Document processing also may include image filtering to remove noise and damage from the record.

Document processing is a precursor to handwriting recognition and optical character recognition (OCR). It is used to find and separate zones of handwritten text from zones of printed text and it also identifies columns, all in preparation for passing parts of the image to the handwriting recognition or OCR software. Along with the sub-images, the document processor can also pass contextual information to the recognizers. In this way, the recognizers can improve the quality of their results by taking advantage of the formatting and spatial information extracted from the document by the document processor. This information can give clues as to the recognizers as to whether the text may be names, dates, places, or relationships, etc.

The text transcripts produced and output from a recognizer are returned back to the document processor where they are collated together along with spatial and contextual information into an annotated transcript of the document. The resulting transcript is a faithful copy of the document that preserves the characteristics of the original as much as possible, making it useful for further analysis and research.

Handwriting Recognition and Optical Character Recognition

Recognition is the technology to extract text from images. Optical character recognition (OCR) is a less difficult problem (albeit still very difficult) than handwriting recognition. Because of that fact and because it has many commercial applications, OCR is much more advanced than handwriting recognition. There have been a number of research projects related to handwriting recognition that have been attempted. Word spotting technology separates text into words and then groups those words into clusters where the words in each cluster are all instances of the same word. Automated record indexing attempts to only extract names from a record in order to build an index to help people to quickly locate a record that might be of interest. And finally, an automated record transcription system attempts to make a full transcription of a record into text. Dictionaries, domain experts, and natural language processing systems can help recognizers choose amongst the set plausible transcriptions in order to choose the best transcription that makes the most sense.

Document Markup

Document markup (also called segmentation and classification) is a technology to categorize text fragments in a document into semantic elements (fields) such as names, dates, places, relationships, etc. After fragments have been categorized, tags or markup can be added to the transcript of the document to convey this information. Document markup adds semantic understanding to text that can be used by other computer programs. A document that has been marked up has improved search results, simplified data extraction, and can make better use of digital experts. RDF, OWL, and Microdata are semantic web standards used for adding tags and metadata markup to a document transcript. Schemas for record markup must support data extraction, which aggregates semantic elements into facts (or assertions) that can be used to establish the identities of individual persons and their relationships to each other.

Data Extraction

Data extraction (also called normalization) extracts structure from an otherwise unstructured text by aggregating marked semantic elements together into facts. This is done by an expert system that analyzes the marked-up and annotated transcripts of records to infer genealogical facts and relationships. These facts and relationships are then represented in a normalized form that can be searched and manipulated. The normalization process often requires using date and place authorities in order to standardize dates and locations.

Family-Linked Data Models

A data model is a structured and normal way of representing data in a regular and well-defined fashion. It could be a set of tables in a database or a set of classes in an object oriented language. Whatever its form, it must be able to represent the full variety of data for which it was designed.

Family-linked data models must represent all types of family relationships, which include parent-child relationships and spousal relationships. The data model must also be able to represent all types of genealogical facts and preserve links between those facts and the source records from which the facts originated. The links between facts and source records are kept to preserve provenance and to be able to substantiate conclusions based on those facts. And lastly, facts must be able to be linked to the people or relationships to which those facts refer.

Most of these links are many-to-many. Multiple source records can substantiate a single fact and a single source record can contain many facts. Facts can refer to multiple people and relationships and an individual person can be described by multiple facts. Many commercial data models cannot correctly represent these many-to-many links. There is a tradeoff to be made between the complexity of the data model and its usefulness. If a model tries to be too fancy, it is doomed because few people will be able to understand it or be able work with it.

Record Linking and Record Merging

Both record linking and merging require a good solid data model to support those operations, as discussed above. Record linking is technology to determine when two or more different records refer to the same individual. When records are linked, then facts derived from those records can all be linked to the same individual. Record merging is similar to record linking. Record merging is when two individual records are determined to be referring to the self-same individual. An individual is described by first, a collection of facts (assertions) that are ordered by time to build a timeline for that person, and second, a collection of relationship links to other persons such as parents, spouses, and children. When two records are merged, then the facts and relationships in both records are combined together into one record. Duplicate facts and relationships are combined into one. Non-duplicate facts and relationships are both added to the newly merged individual record. Ideally, at the end of the merging process, the merged individual should not have any conflicts. To better support family reconstitution, it might be useful if merges could be "undone" to allow for backtracking in the family reconstitution algorithm.

Record matching can be graded by two standard measures. *Precision* is the percent of matches found that are actually correct, and *recall* is the percent of correct possible matches that were actually found. Ideally, a matching scheme is 100% precise (no incorrect matches made) with 100% recall (all possible correct matches were made). Matching is most frequently done by some sort of scoring heuristic which uses a minimum threshold to determine if the candidate match is counted as a true match or not. There is a tradeoff to be made in choosing the threshold. If it is too high, then correct matches are not made and recall goes down. If the threshold is too low, then incorrect matches are made and precision goes down. It is usually impossible to choose a threshold that gives perfect scores in both precision and recall.

Digital Experts and Assistants

Digital experts are specialized domain experts that generally have large knowledge bases to which they can refer. Digital experts can work for a human, through a computer-human interface, in which case they are called digital assistants. A digital assistant must be able to understand and answer simple questions about its domain. A digital expert can also work with other computer programs through an API. As an expert, it should be able to analyze a piece of data or a hypothesis and given an educated opinion about it, which can come in the form of a probability of correctness derived from its knowledge base.

For genealogy, there are a number of useful digital experts:

Name authority A name authority uses a knowledge base of all names and how common they are at a particular time and place. It can also say whether two names could be used for the same person. (Mary == Polly?)

Place authority A place authority uses a knowledge base of all places, their names, their locations, the date range they existed, their origin, their jurisdictions, the jurisdictions that were contained within them, and the jurisdictions in which they were contained. They are like a computer gazetteer. They can answer questions such as "what is the distance between two places?" and "what is the likelihood that a person moved from one place to another?".

Date authority A date authority does date computations and converts dates from non-standard forms (like Feast days) to the standard calendar.

Record authority A record authority uses a knowledge base of what source records exist for given jurisdictions and what information they contain. It can suggest what records to search to find additional information about a given person.

Timeline (chronology) authority A timeline authority can analyze a timeline for a specific person and estimate likelihoods of correctness based on statistical analysis of similar communities and probability distributions of ages for specific types of events. It can estimate ages and dates based on greatest likelihoods.

Population Reconstitution

Population reconstitution is to reconstitute all of the families for a complete population living in a given locality by merging information from all sources for that locality (or as many sources as are available). Population reconstitution is valuable for demographers, geneticists, and genealogists alike.

Populations are reconstituted using probabilistic record linking and tree matching algorithms to merge information from many different sources and record types such as censuses, civil records, directories and church records. It is most often done for an entire community in a fixed locality because that is the way the records are kept and organized. It is premised on the assumption that when the complete set of records for a given location is analyzed, it allows you to make conclusions based on the absence of contradictory evidence as well as direct evidence. Spock stated it this way: "If you eliminate the impossible, whatever remains, however improbable, must be the truth." With all of the available information, the task then is to find a population that best fits all of the available data. People moving in and out of the community further complicate this difficult task.

This module comes at the end of the long module sequence that digitizes, analyzes, transcribes, marks-up, and extracts data from documents, which data is normalized into a model, matched, and merged. The problem with current population reconstitution studies is that the source data must be prepared by hand which takes a tremendous effort, and it is rare to have all of the documents for a locality extracted and put into a common data model. In our utopic system, the computer can do all of that work.

Intelligent Search, Visualization, and Human Computer Interface

Once data has been extracted and organized into families, it must be made searchable for people to make connections to it. The intelligent search problem is to search through a large collection of records to find records, persons or facts that are consistent with what is known about a given person or family. The search problem is related to the record-linking problem that has already been discussed. On first glance, it doesn't appear that it should be a difficult problem, but it is harder than it looks and it seems that no one can get it right. You have all worked with genealogical search engines that return results that cannot be correct and are often absurd (low precision) or fails to return correct results or even results that are plausible candidates (low recall). An example of an intelligent search request is to ask for all possible children of a given couple in your database. The search engine would need to put together search criteria obtained from your data to start the search. The search engine should be able to operate in presence of small errors in the data and return persons that have high likelihoods of actually being the children that you want. Absurd results like proposing children born 50 years after the death of the parents should never be returned in the search results. However, instead of computing likelihoods, most search engines score matches heuristically and then use a threshold to determine whether or not a record "matches" the search criteria.

In general, the way a human being views and interacts with genealogic data is an on-going topic of research of interest to our community.

Summary

In this paper, I surveyed some of the problems that I view as being on the critical path to a revolutionary break-through in the way we do genealogical research. They are all difficult problems that have had research work done to solve them. However, they still remain unsolved. I do not claim that this list is complete. There are undoubtedly other important problems that impede family history research that I have not surveyed. There may be other important problems that have not even been identified yet.

The Family History Technology Workshop is held each year to discuss these kinds of problems with the hope that technology can put the ultimate goal of assembling a family tree for all people within reach.

References

Document Digitization and Preservation

Barry Lunt, Robert Davis, Matthew Linford, Family History Archives: Research on Permanent Data Storage", FHTW'2012.

Oliver Nina, Roger Pack, "High Dynamic Range Imaging (HDRI) for Preservation of Historical Documents", FHTW'2011.

Michiel Anderson, Oliver Nina, Michael Wynn, "Contrast Enhancement using Locally Adaptive Binarization Techniques", FHTW'2011.

Barry Lunt, Ryan Sydenham, Feng Zhang, Matthew Linford, "Digital Data Preservation: The Millennium CD and Graceful Degredation", FHTW'2007.

Gene A. Ware, Doug M. Chabries, Richard W. Christiansen, Curtis E. Martin, "Multispectral document enhancement: ancient carbonized scrolls", Proceedings of the 2000 Geoscience and Remote Sensing Symposium, vol. 6, pp. 2486 – 2488, July 2000.

Gene A. Ware, Doug M. Chabries, Richard W. Christiansen, James E. Brady, Curtis E. Martin, "Multispectral Analysis of Ancient Maya Pigments: Implications for the Naj Tunich Corpus", Proceedings of the 2000 International Geoscience and Remote Sensing Symposium, vol. 6, pp. 2489-2491, July 2000.

Document Processing

Aaron Stewart and David Embly, "Extraction Names Using Layout Clues: An Initial Report", FHTW'10, pp 49-54.

Xujun Peng, Huaigu Cao, Rohit Prasad, Krishna Subramanian, Prem Natarajan, "An Efficient Method for Extracting Family Records from Mixed-type Forms", FHTW'2012.

Krishna Subramanian, Huaigu Cao, Xujun Peng, Rohit Prasad, Prem Natarajan, "Image Registration and Text Recognition for Structured Census Documents", FHTW'2012.

Oliver Nina, Bryan Morse, "Interactive Smoothing of Handwritten Text Images Using a Bilateral Filter", FHTW'2009.

Oliver Nina, William Barrett, "Thresholding of Text Documents Using Recursion of the Otsu Algorithm", FHTW'2007.

Douglas Kennard, William Barrett, "Progress with Searchable Indexes for Handwritten Documents", FHTW'2007.

Handwriting and Optical Character recognition

Kevin Bauer, “Intelligent Pen: A Least-Cost Search for Tracing of Handwriting”, FHTW’2014.

Robert Clawson, William Barrett, “Intelligent Indexing: A Semi-Automated, Trainable System for Field Labeling”, FHTW’2014.

Xuhun Peng, Elizabeth Boschee, Huaigu Cao, Rohit Prasad, Krishna Subramanian, Prem Natarajan, “Information Extracton from Historical Semi-Structured Handwritten Documents”, FHTW’2012.

Douglas Kennard, William Barrett, Thomas Sederberg, “Handwriting Recognition (HR) of Family History Documents using a 2-D Warping-based Word-level HR Approach”, FHTW’2012.

Robert Clawson, William Barrett, “Extraction of Handwriting in Tabular Document Images”, FHTW’2012.

Douglas Kennard, “Word-Spotting for Automatic Tag Suggestion in the BYU Historic Journals Project”, FHTW’2009.

Thomas Packer, Oliver Nina, Ilya Raykhek, “Using a Hidden-Markov Model in Semi-Automatic Indexing of Historical Handwritten Records”, FHTW’2009.

“Word spotting”, http://ciir.cs.umass.edu/irdemo/hw-demo/wordspot_retr.html

Document Markup

Robert Gardner, Dave Barney, “HTML5 Microdata Markup for Genealogy Sites”, FHTW’2013.

Robert Gardner, Tony Ruscoe, Dave Barney, “Genealogy, Microdata, and Search Engines”, FHTW’2012.

Carl Christensen, Deryle Lonsdale, “Domain-Independent Data Extraction: Person Names”, FHTW’2009.

David Wiggins, Scott Woodfield, “Name, Date, and Place Recognition in Unstructured Text”, FHTW’2009.

Jonathan Baker, Hilton Campbell, Jordan Crabtree, David Embley, “Pattern Markup Language: A Pattern-Based Tool for Quickly Automating Genealogy Data Extraction”, FHTW’2008.

Jay Askren, “The Semantic Web for Family History”, http://jay.askren.net/Projects/SemWeb/

Sue Adams, “Claverley Property Document Analysis”, Part I – Transcript http://familyfolklore.wordpress.com/2013/10/14/claverley-property-document-analysis-part-1-transcript/, Part II – Semantic Markup http://familyfolklore.wordpress.com/2013/10/21/claverley-property-document-analysis-part-2-semantic-mark-up/, Part III – Places

http://familyfolklore.wordpress.com/2013/11/05/claverley-property-document-analysis-part-3-places/, Part IV – People and Identity http://familyfolklore.wordpress.com/2014/01/15/claverley-property-document-analysis-part-4-people-and-identity/, Part V – Next Steps http://familyfolklore.wordpress.com/2014/02/01/claverley-property-document-analysis-part-5-next-steps/, 2013-2014.

Data Extraction

Peter Lindes, Deryle W. Lonsdale, and David W. Embley, “Finding Genealogy Facts with Linguistic Analysis” FHTW’2014, March 2014.

Peter Lindes, Deryle Lonsdale, David W. Embley, “OntoSoar: Using Language to Find Genealogy Facts”, FHTW’2013, March 2013.

Joseph Park, David Embley, “Extracting and Organizing Facts of Interest from OCRed Historical Documents”, FHTW’2013.

Deryle Lonsdale, David Embley, Stephen Liddle, Joseph Park, “Extracting Information from French Obituaries:, FHTW’2012.

Packer, Lutes, Stewart, Embley, Ringer, Seppi, and Jensen, “Extracting Person Names from Diverse and Noisy OCR Text”, FHTW’10, pp. 55-57.

Charla Woodbury, David Embley, and Stephen Liddle, “Automatic Extraction From and Reasoning About Genealogical Records: A Prototype”, FHTW’10, pp. 59-76.

Andrew McCallum, Karl Schultz, and Sameer Singh: FACTORIE: Probabilistic Programming via Imperatively Defined Factor Graphs, NIPS'09

Andrew McCallum, “Information Extraction: Distilling Structured Data from Unstructured Text “, *ACM Queue*, Volume 3, Number 9, November 2005.

Family-Linked Data Models

David W. Embley, Scott N. Woodfield, “A Superstructure for Organizing Family History Information”, FHTW’2014, March 2014.

Joan Campanya Ares, Jordi Conesa Caralt, Enric Mayol Sarroca, “Modeling Genealogical Domain: An Open Problem”, KEOD 2012, Barcelona.

Record Linking and Record Merging

Patrick Schone, “Development of an Evaluation Paradigm for “RecordMatch” and its Application to GenMergeDB Clustering Results”, FHTW’2011.

D. Randall Wilson, “Genealogical Record Linkage: Features for Automated Person Matching”, FHTW’2011.

David S. Barss, “Using a Lineage Linked Family Perspective over Household to Improve Record Linkage Success with Census and Other Data Collections”, FHTW’2010, pp. 19-32.

Stephen Ivie, Yao Huang Lin, Christophe Giraud-Carrier, “Utilizing Stacking for Feature Reduction in Graph-Based Genealogical Record Linkage”, FHTW’2008.

Randy Wilson, “Genealogical Record Linkage on International Data”, FHTW’2008.

Steve Ivie, Graham Henry, Haven Gatrell, and Christophe Giraud-Carrier, *“*A Metric-based, Machine Learning Approach to Genealogical Record Linkage”, FHTW’2007.

Thomas W. Jones, “Merging Identities Properly, Jonathan Tucker Demonstrates the Technique”, NGS Quarterly 88, pp. 111-121, June 2000.

Digital Experts & Assistants

Randy Wilson, “Date Range Propagation in Genealogical Databases”, FHTW’2012.

Justin Seliger, “The Folk Date Pattern”, FHTW’2011.

Patrick Schone, Stuart Davey, “A Multilingual Personal Name Treebank to Assist Genealogical Name Processing”, FHTW’2012.

Population Reconstitution

Sue Dintelman and Tim Maness, “Reconstituting the Population of a Small European Town using Probabilistic Record Linking: A Case Study”, FHTW’2010, pp. 33-44.

Randy Wilson, “Can a Layered Approach to Historical Family Reconstitution Make a Difference in the Final Results?”, FHTW’2008.

Intelligent Search, Visualization, and Human Computer Interface

Ryan Cheatham, Curtis Wigington, William Barrett, “Virtual Pedigree– Genealogy Without Borders”, FHTW’2014.

Robert Ball, David Cook, “A Family-Centric Genealogy Visualization Paradigm”, FHTW’2014.

Bill Harten, “Introduction to Puzilla.org: A Condensed Descendants Viewer for FamilyTree”, FHTW’2014.

Geoffrey Draper, Richard Riesenfeld, “Interactive Fan Charts: A Space-saving Technique for Genealogical Graph Exploration”, FHTW’2008.

Mark Tucker, “10 Things Genealogy Software Should Do”, FHTW’2008.

Barry Lunt, Ryan Sydenham, Feng Zhang, Matthew Linford, “Digital Data Preservation: The Millennium CD and Graceful Degredation”, FHTW’2007.