# We Have Hinting – What's Next?

**David Ouimette, CG, CGL**
david.ouimette@familysearch.org

## What Is Family Reconstitution?

Family reconstitution is the process of identifying individuals, families, and lineages across a large historical population through analysis and correlation of evidence gleaned from broad-coverage records rich in family history details. While traditional family history research focuses on one family at a time, family reconstitution examines all the evidence on all families simultaneously and produces a comprehensive set of family trees for an entire locality. Researchers and organizations apply automated and manual techniques to reconstitute families within a parish or province.

## Why Do Family Reconstitution? Help More People Find More Ancestors Faster.

FamilySearch and other genealogical organizations help more people find more ancestors faster by developing technologies to increase access to historical records and the details they reveal about ancestors. Moving up the record-access continuum, more people have more success finding more ancestors faster:

Family reconstitution is a natural next step beyond existing hinting technologies. Whereas record hints introduce individual record matches in online trees, family reconstitution can introduce well-documented family groups as matches in trees.

## Successful Family Reconstitution – Necessary Conditions

What conditions must exist in an historical population and its corresponding record collections so that family reconstitution will produce accurate and comprehensive results?

- **Non-migratory population**
  - Families persisted in the same place for multiple generations
  - Feudal, agrarian, parish communities
  - Tenant farmers in pre-industrial communities
  - Serfs tied to the land
- **Record coverage**
  - Broad population coverage of surviving records
  - Catholic or Protestant parish registers that document vital events comprehensively
  - Civil registers of births, marriages, and deaths during time periods of full compliance
  - Nominal census population schedules, identifying all individuals in all families
- **Richness of information**
  - Records must identify individuals uniquely and place them within families

 © 2016 David Ouimette, CG, CGL

- Indexing projects must capture personal and family details found in the best records
  - Rich identifying information (e.g., maiden names of women)
  - Rich relationship information (e.g., parents of all principle individuals)
  - Rich family information (e.g., lineages, house numbers)

Software algorithms may perform well reconstructing isolated, non-migratory populations that are well documented in historical parish registers. Successful examples include:

- *Programme de recherche en démographie historique* – aka PRDH, the Research Program in Historical Demography at the University of Montreal, has reconstructed all families and lineages in French Canada from the 1600s to 1800
- *Ortssippenbücher*, lineage books that document all families within a German community
- Various parish reconstitution projects in the United Kingdom

## The Family Reconstitution Process

1. **Select and index records** – focus on record coverage and richness of information
2. **Standardize data** – improve precision and recall by handling name variations optimally
3. **Block and sort records** – reduce computational complexity
4. **Compare and cluster records** – score and bin possible matches
5. **Coalesce individuals and families** – draw genealogical conclusions

## Step 1: Select and Index Records (focus on record coverage and richness of information)

Choose record collections with broad coverage such as parish registers, civil registers, and nominal census population schedules. Select indexing fields necessary to identify individuals uniquely and place them within their families: names, event dates, event places, relationships, ages, residences, occupations, and any other fields that could help discriminate between individuals. Isolate personal name components into separate fields. Use culture-specific standardization tables when indexing given names and family names. Use gazetteers to ensure accurate indexing of places.

## Step 2: Standardize Data (improve precision and recall by handling name variations optimally)

Precision and recall improve dramatically when appropriate standardization techniques cleanse the indexed data prior to blocking and record comparison. All data may contain noise in the form of spelling variations, incorrect application of diacritics, various delimiters, indexing errors, and errors in the original records. In addition, each indexed field suffers from problems unique to its record type.

**Given names** introduce many challenges: spelling variations, nicknames, multiple given names per person, name changes over a lifetime, translation into difference languages, and other cultural issues.

**Surnames** have their own challenges: spelling variations, cultural naming patterns (e.g., French Canadian *dit* names, Swedish patronymics, Norwegian farm names, Mexican dual surnames), surname declensions, name changes over a lifetime, and other issues.

While a number of surname-encoding algorithms exist, they only attempt to mitigate spelling variants and even then only find applicability in limited contexts. Some of the more popular encodings are:

- Russell Soundex (NARA, Census)
- Daitch-Mokotoff Soundex (Slavic, Yiddish)
- NYSIIS (New York)
- Oxford Name Compression (combines NYSIIS and Russell Soundex)
- Double-Metaphone (European and Asian)

 © 2016 David Ouimette, CG, CGL

Machine learning, based on a substantial training database, can also help manage spelling variations.

Ideally, name standardization employs culture-specific name tables to manage variants like nicknames that are not detectable through algorithmic distance metrics nor machine learning. For instance, the PRDH database has a comprehensive table of surname variants and a mapping of all *dit* names to their corresponding surnames, tabulated from three hundred years of Catholic parish registers.

**Localities** need to be standardized, geocoded, and associated within all relevant geopolitical and ecclesiastical hierarchies. This is especially true if localities are used as a blocking key in the next step.

Other fields such as age, residence, and occupation also merit proper formatting and standardization.

## Step 3: Block and Sort Records (reduce computational complexity)

In a large database, exhaustive pairwise comparison of records is costly and wasteful, as most records are unrelated. Yet records need to be compared to identify which pertain to the same individuals and families. It's best to partition a database into small groups of records that likely contain information on the same individuals. Blocking involves splitting a database into groups of records for purposes of record comparison. "Blocking or sorting keys that are required for all indexing techniques … are traditionally being defined manually by data matching and domain experts." (P. Christen, *Data Matching*, p. 97)

Which blocking keys provide the best partition of a large database of, say, parish register indexes into useful blocks for record comparison?

- Geocoded locality (e.g., parish, village, or region)
- Standardized surname, surname encoding, or first letter of surname
- Combination of locality and name keys

The decision on blocking and sorting keys considers computational economy, geographic movement of families over generations, and the stability of family names over time. If the blocking scheme partitions the database into disjoint blocks, such as by parish or by first letter of surname, each record appears in only one block. If the blocking strategy uses a sliding-window or sorted-neighborhood approach, such as by clusters of adjacent parishes or by sorted names, a record may appear in more than one block.

## Step 4: Compare and Cluster Records (score and bin possible matches)

Each baptism, marriage, death, or census record has names, relationships, and event dates and places that should be indexed for record-to-record comparison. Pairwise record comparison produces a distance vector comprised of field-comparisons scores weighting factors. Distant metrics and weighting factors may be tuned for each field *a priori* or via machine learning. Adjustments are made based on culture, name variety, and index quality. Pairwise comparison of data, or a running comparison of sorted data, yields candidate individuals and family groups for refinement in the next step.

When marriages identify all four parents by their birth names, marriage records can build lineages directly, linking all married people to their parents for as many generations as there are records. If name standardization adequately manages name issues and if the blocking strategy allows for sufficient breadth of record comparison, these marriage linkages can provide a solid skeleton for reconstituting all family groups within the target population. Subsequent comparisons of birth and death records via the names of parents and spouses can complete the reconstitution project.

Whenever baptisms, marriages, or burials lack full birth names of the father and mother of the principal individuals, this compare-and-cluster stage seeks to bin similar records into candidate family groups for further analysis and correlation, including multi-record comparison and analysis of indirect evidence.

 © 2016 David Ouimette, CG, CGL

## Step 5: Coalesce Individuals and Families (draw genealogical conclusions)

This final step uses direct and indirect evidence—from pairwise and multi-record comparisons, respectively—to compile information into individuals and families. Genealogical conclusions made in family reconstitution should follow the Genealogical Proof Standard, including these five components (quoted from BCG's *Genealogy Standards*, 50th-Anniversary Edition (2014), pp. 2-3):

- Reasonably exhaustive research
- Complete and accurate source citations
- Critical tests of relevant evidence through processes of analysis and correlation
- Resolution of conflicting evidence
- Soundly reasoned, coherently written conclusion

The previous step produced record clusters defining candidate individuals and families. In the simpler cases, direct evidence has definitively matched records to produce individuals and assign family relationships. In more complex cases, candidate individuals and families need additional analysis and correlation of indirect evidence in order to finalize decisions on which records merit combination into individuals and families. Multi-record comparisons seek answers to these questions:

- Do the births of candidate children conflict?
- Does the age of the mother conflict with birth dates of candidate children?
- Does the death of the spouse precede the subsequent marriage of the widow(er)?
- Do surviving children of the same name live together?
- Does the record or individual fit equally well into more than one family?

When there are multiple conflicts, it may be necessary to apply a best-fit algorithm (such as passing a cost matrix through the Hungarian algorithm). The more complex the situation, the more likely the need for expert manual review. In certain boundary cases—such as earlier registers which lack key information fields—it may be difficult or impossible to reconstitute families with or without expert review.

## Bibliography


- Board for Certification of Genealogists. *Genealogy Standards*, 50th-anniversary edition, Nashville: Ancestry, 2014.
- Christen, Peter. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Berlin: Springer, 2012.
- Fure, Eli. "Interactive Record Linkage: The Cumulative Construction of Life Courses," Demographic Research, vol. 3, no. 11, 2000.
- Quass, Dallan, and Starkey, Paul. "Record Linkage for Genealogical Databases," *ACM SIGKDD '03 Workshop on Data Cleaning, Record Linkage, and Object Consolidation,* August 24-27, 2003, Washington, D.C.
- Vézina H., Dillon L., Bellavance C. "Quebec population resources: Towards an integrated infrastructure of historical microdata (1621-1965)." Presentation at the Congrès de l'UIESP, Busan, Corée, 26-31 Aug 2013.
- Wilson, D.R. "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage", Proceedings of the 2011 Joint International Conference on Neural Networks, pp. 9-14, 2011.
- Wilson, D. Randall. "Genealogical Record Linkage: Features for Automated Person Matching," *RootsTech 2011*, February 4, 2011, Salt Lake City, Utah.


 © 2016 David Ouimette, CG, CGL