Automatic Merging of Pedigree Information

Annual Workshop on Family History Technology
April 3, 2003

Sue Dintelman and Tim Maness
Pleiades Software Development, Inc.
Source of Duplicates

- **Common Ancestry Trees**
  - Most large pedigrees have branches that intermarry

- **Combining Data Sources**
  - Working with other family members to build a common genealogy
  - Utilizing on-line or other sources to expand your genealogy
Current Solutions

- Not automated
- Utilize limited clustering options
- Utilize limited family information (Parents’ names)
Goals for Merge Utility

- Automatic
- Fast
- Accurate
- Eliminate duplicates in a single family database
- Combine multiple family databases
Record Linking Background

- Decide if two records are for the same individual
- Use sum of weights for a comparison of each common field in the records
- Use a cut off score to choose “true” links
Sample Scores
Problems Linking Individuals in Family Data

- Few fields that can actually be compared (name, birth date and place, death date and place)
- Many names will be similar or identical because of naming conventions
- Many places will be the same because these are families
Advantages Linking Individuals in Family Data

- Family members provide additional field values for comparison

- Additional family information helps prevent incorrect matches
Other Record Linking Considerations

- Misspellings of names and places
- Incorrect dates
- Initial inconsistencies
  - Any family database with 20+ generations has some type serious inconsistency
The Process

- Data preparation
- Find initial duplicates
- Use a recursive process to find other duplicates
Data Source Preparation

- Find loops (an individual is his own ancestor)
- Find inconsistent information (a person is born before his parents)
- Identify connected components
- Pre-process names, places and dates
Generate Duplicate List

- Cluster using last name variation
  - Transducer
- Compute score
  - Individual component
  - Family component
- Choose the links with the highest scores
Merge Duplicates

- For each pair of duplicates:
  - Combine data
  - Recursively consider the relatives of the duplicates

- Add any new duplicates to the list
New Duplicate

Misspelling:
- Jones, Jerrolyn, Mary
- Jonesanderson, Jerrolyn, Mary

Duplicate sib:
- Kimball, Lanette 3/4/1905
- Kimball, Lannette 0/0/1905
The Merge Reports

- List of people who merged
- List of new people
- List of parent problems
Example Parent Problem

Jonathan Anderson, born 07/07/1848 Nauvoo, Hancock, OH
  Spouse: Maria Babcock, born 08/09/1852 Nauvoo, Hancock, OH
  (five children Ann, John, Alex, Samantha, Elizabeth)
  Mother: Emily Adams, born 02/19/1823 Pomphret, Chautauqua, NY

Father: Jonathan P. Anderson, born 10/28/1824 Wartrace Creek, Bedford, TN

Jonathan Anderson, born 07/07/1848 Nauvoo, Hancock, OH
  Spouse: Maria Babcock, born 08/09/1852 Nauvoo, Hancock, OH
  (five children Ann, John, Alex, Samantha, Elizabeth)
  Mother: Theresa Johnson, born 04/17/1825 New York City, NY
  Father: Jonathan K. Anderson, born 08/15/1820 Weakly, TN
GenMerge

- Automates finding and eliminating duplicates in a single data source or when combining data sources
- Fast
- Accurate
- Allow review of inconsistencies