Graph-Based Remerging of Genealogical Databases

D. Randall Wilson

fonix Corporation
Draper, Utah, USA

e-mail: WilsonR@fonix.com
or randy@axon.cs.byu.edu

Workshop on Technology for Family History and Genealogical Research
Brigham Young University
March 29, 2001
“Remerging” Problem

Original Database

Share a copy

Both make independent updates....

Now what??
Common Approaches

• Give up
  • One person does everything, and everyone else is uninvolved; or
  • Everyone duplicates work for themselves.

• Visual Inspection, and hand-typing

• Unix “diff” command, and hand-typing

• Match/Merge function
  • Import second database into first
  • Decide which pairs of similar people should be merged back together

Time wasters :(
Better Solutions

• Locking
  • One person has *master* database
  • Others can “check out” portions
    *[but overly restrictive]*

• Unique ID Numbers
  • Program assigns unique ID numbers
  • ID numbers allow automatic match/merging of identical people.
  • *[but ID numbers may not survive translations to/from other software]*

• Graph-Based Merging Algorithm
Graph-Based Merging

• No need to check out (lock) portions of the database.

• No need for ID numbers

• No need to examine people who have not changed.

• Retroactive: Works on databases that have already diverged.
Merging Algorithm

I. Sort both databases
   • Surname, given name
   • Birth date, birth place
   • Death date, death place
   • ID numbers, if available

II. Find “matching” person
   • Search lists in parallel; O(N+M) time.
   • Find people with same personal information
   • Then search relationship graph
Merging Algorithm (cont’d)

Search relationship graph

child 1
child 2
child 3

spouse

individual

father

mother

child 1
child 2
child 3

spouse

individual

father

mother

Graph-Based Remerging
Merging Algorithm (cont’d)

Labeling subgraphs

Continue

Graph-Based Remerging
Merging Algorithm (cont’d)

III. Choose largest subgraph

IV. Incorporate new information
   • Additional individuals

V. Connect subgraphs.
   Continue until all incoming information has been included or rejected.

• Additional information
• Conflicting information
• [Missing information]
Uses for Graph-Based Merging

• Collaboration with family members
  • Independent updates/work/research
  • Collect information on immediate family

• Family history organization
  • Archivist assigns work to helpers
  • Research director, archivist, helpers all add to database concurrently.

• Database on multiple computers
  • Desktop/laptop; home machine; etc.

• Include previously excluded info

• Find differences between databases
Advantages of using graph-based merging for remerging genealogy databases

• Much easier than manual approaches
• Much faster than global match/merge
• No need for checking out (locking)
• No need for ID#s
• Not restricted to single platform or software package
• Retroactive solution
• User controls changes to their data
Further Work

• Actual implementation

• Identifying “similar” people
  (to distinguish between additional individuals vs. additional or conflicting information)

• Note-merging
  • Reordered notes
  • Minor changes vs. new notes

• Multimedia

• Global differences/Style
  • “Lee Co., VA” vs. “,Lee,VA”
  • Surname capitalization

• Remembering decisions
  • Avoid repeating same decisions next time.