Efficient Genealogy through Personal Extraction and Automated Verification

D. Randall Wilson

WilsonR@fonix.com

Overall Goal

- Create a global genealogical database
- Avoid duplication of effort
- Allow true collaboration
- Make efficient use of people's time

Outline

- Extraction work
- Massive personalized extraction
- Evidence-based genealogy databases
- Automated Verification

Traditional Genealogical Research

- Search for references to relatives
  - Libraries (books, microfilms)
  - Court houses
  - Internet (GEDCOM, genealogy sites)
- Extract relevant information
- Enter into database

Traditional Research

- Most of time spent searching
- Spend long time per name
- Pass over many others
- Hit and miss
- Duplication of effort (“re-search”)

Extraction:
More Efficient

- **Records are consecutive**
- **Learn once/enter many (assembly line)**
- **Source citation automatic**

Later searching much faster

Extraction:
Other Advantages

- **Simpler training**
- **Straightforward**
- **Complete coverage**
- **Avoids duplication**

Personal Extraction

- Let people choose what to extract.

  => more interesting.
- Allow extraction over the internet.
  - Massive participation
  - Download images
  - Do work on-line or off-line
  - Upload extracted data

Extracted Record

- **Represents a *reference* to a person.**
  - "Identity" (Harten, 2002).
  - "Persona" (GENTECH, 2002).
- **Contains information obvious from document itself.**
- **Has relationships to others in document.**

James, the son of William Ball.

Fundamental Inefficiency of Current Databases

- "Deal only with tidy, final conclusions."
- "Sidestep harder job of storing, linking and sharing evidence…and of representing conflicting and changing opinions."
- "Offer no basis for future evidence evaluation." (Harten, 2002).
- Weak connection to original sources.

Evidence-based Database: Linking Extracted Records

- Extracted records are basic building blocks.
- *Person* record built by linking extracted records.
- *Best view* automatically built from records.
- Trace conclusions back to original sources.

Managing Conflicts

- **Automatic priority.**
  - Earlier, primary, more reliable documents.
- **Manual priority.**
  - Use proper justification.

Justification
(Logic, Reasoning)

- **Linking records together**
  - Similar names, places, dates, relationships
- ***Not* linking records together**
- **Transforming data**
- **Preferring record in "best view"**
- **Making logical conclusion given data.**

Verification

National Genealogical Society:

"Family history researchers—

- *Record the source for each item of information…*
- *Seek original records…as the basis for their research conclusions.*
- *Use compilations…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."*

Automatic Verification

- **Computerized verification**
  - Rigid justification
  - Gather statistical data
  - Automated justification, linking, etc.
- **Signature-based verification**
  - Easier in short term
  - Human verification with username “signatures”
  - User accounts with reliability ratings
  - Computer algorithms with IDs and ratings

Verify Each Stage

- Document→[Microfilm]→Image
- Transcription (if any)
- Extracted records
- Data normalization
- Links, assertions, conclusions
- Justification (logic, reasoning)

Summary

- Extraction first, then search ⇒ Faster
- Massive personal extraction ⇒ Participation
- Extracted records as basic building blocks
  ⇒ Connect to original sources
- Justification ⇒Preserve reasoning.
- Verification ⇒Trust each other's work.

Conclusions

- True collaboration.
- Avoid duplication.
- Eventual complete coverage.

WilsonR@fonix.com