# Bidirectional Source Linking: *Doing Genealogy Once and For All*

Randy 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

April 4, 2002

# Introduction

Suppose our goal was
to build the most complete
genealogical database possible
from all of the records on Earth.

- How long would it take?
- How would we know when we were “done”?

Much time spent on duplication of efforts.

# Transcription Example

100,000 volunteers
1,000,000 microfilms

If done right:

- Each record transcribed once
- Done when all microfilms have a completed transcription.
- Transcriptions available for fast searching, etc.

# Transcriptions Gone Bad

100,000 volunteers
1,000,000 microfilms

If not organized:

- Some records transcribed many times, others not at all.
- Never know when we're done.
- Data stored on volunteers' own hard drive, scattered across internet, etc.

# Current Approaches to Genealogical Research

- Search for records
- Write down relevant information
- Enter conclusions into genealogy database
- Share data with relatives, on web site, or in on-line repositories
- Next person who finds same record repeats the process

# Evidence-based Research

*Evidence* vs. *conclusions:*

*John Smith was born in 1835.*

Genealogists often agree on the evidence, but may disagree on the conclusions.

Most software encourages users to enter the conclusions first.

GENTECH Lexicon Working Group (www.gentech.org):

- Enter *evidence first*
- Make *assertions* from this evidence
- Further assertions on top of these
- Use preliminary conclusions to build more advanced conclusions.

GENTECH

# Genealogical Data Model

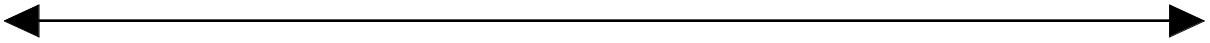
- Shows form of evidence, assertions and conclusions.
- Software should help users to store “the reasoning behind the genealogical conclusions reached, along with all the evidence that led to those conclusions.”
- Exactly one place for each piece of evidence.
- All data keeps an audit trail (even when merged).

# Bidirectional Source Linking

- Every conclusion points to the evidence that led to it

*and*

- Every piece of evidence points to the conclusions derived from it.



Follow links from original records to:

- Other related records
- Assertions made about information in the record
- Databases that reference the record.

# Doing genealogy "once" and "for all"

1. Find a record

2. Add it to the master library
   - Get a unique, permanent ID
   - Enter its description

3. Scan
   - Preserve image
   - Make image accessible

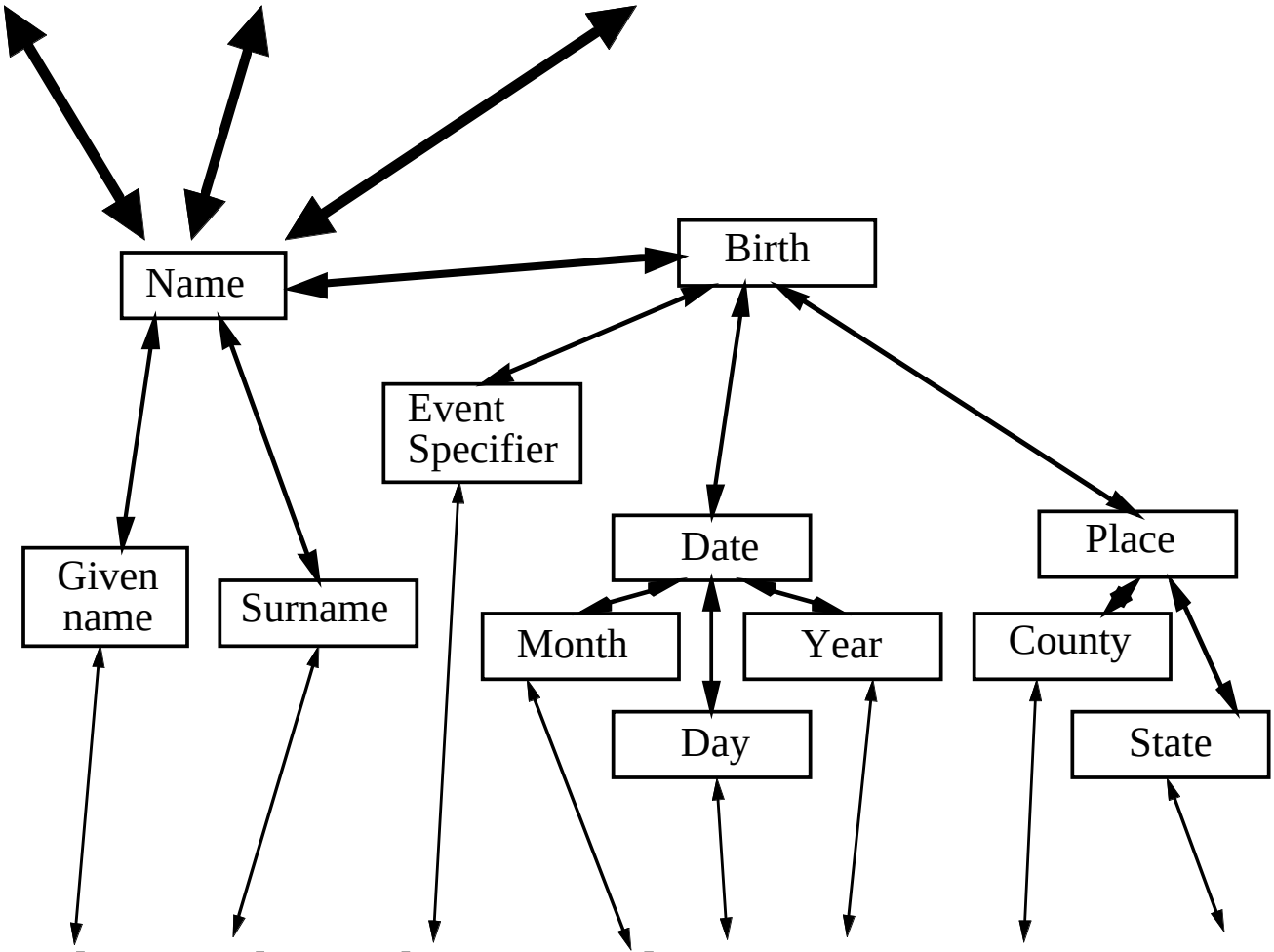
4. Transcribe
   - Make searchable
   - Prepare for tagging

5. Extract structure (tag...)

# 5. Extract structure (tagging)

***John Smith was born July 12, 1835, in Lee Co., VA.***

<NAME ID=Name1, REF-FROM Assert1>
  <GIVEN-NAME>***John***</GIVEN-NAME>
  <SURNAME>***Smith***</SURNAME>
</NAME>
<ASSERTION ID=Assert1, Name1 HAS Birth1>***was***
</ASSERTION>
<BIRTH ID=Birth1, REF-FROM Assert1>***born***
  <BIRTH-DATE>***on***
    <DATE>
      <MONTH>***July***</MONTH>
      <DATE-DAY>***12***</DATE-DAY>,
      <YEAR>***1835***</YEAR>
    </DATE>
  </BIRTH-DATE>
  <BIRTH-PLACE>***in***
    <COUNTY>
      <COUNTY-NAME>***Lee***</COUNTY-NAME>***Co.***
    </COUNTY>,
    <STATE abbr-for "Virginia" >***VA***</STATE>.
  </BIRTH-PLACE>
</BIRTH>

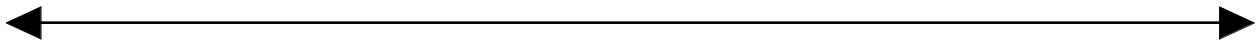



# 6. Link evidence to other records or databases

- Link to (and from) other records that reference the same name
- Link to/from personal database
- Name/place/date authorities

## At every step

- If already done, skip the step.
- If not, do the work once for everyone to use.



## Automated algorithms

- Batch library entries
- Batch scanning
- OCR, handwriting recognition
- Do simple tagging
- Suggest likely links
- Keep audit trail on all information

Do more and more automatically.

# Plenty of work to do

- Define attributes
  - Names, places, dates, etc.
  - Assertions, conclusions
  - Record identification
- General implementation approach
  - XML, GDM, relational DB, etc.
- Resources
  - Storage, bandwidth, etc.
  - Centralized vs. distributed
  - Scanned images for remote work
- Transition to evidence-based approach
- Privacy & copyrights
- Algorithms
  - OCR, handwriting recognition
  - Parsing, tagging, linking, etc.

# Advantages

*Evidence-based approach:*

**Every conclusion shows exactly where it came from.**

- Verification, confidence
- Resolve conflicts
- Meaningful merging:
  share evidence *and* conclusions

*Bidirectional Source Linking:*

**Information in every record shows what references it**

- Avoid duplication of effort
- Know what still needs to be done
- Contribute meaningfully
- Find related records & people
- Avoid using one source to apply to multiple individuals.
- No one permanently overlooked

# Conclusions

- With the right approach, the work could be sped up by orders of magnitude.
- Eventually it should be possible to do genealogy once and for all.