Building Data Models for the Research Process

Dr. Luther A. Tychonievich
Lecturer, University of Virginia
Family History Information Standards Organisation
This talk

• What should we store/share?
  – Claim: research process itself

• How can that be represented logically?
Scope of Family History

- Biological trees
- Familial (and other) relationships
- Evidence and sources
- Attachments (flavor)
- Stories to tie it together

(More on each next)
Simple Ground Truth

• There is a biological ancestry tree
  – Binary going up
  – *n*ary going down

• We do trees well
Real Ground Truth

• Relationships were complicated
  – Adoptions, step-parents, disowning, foster homes, switched-at-birth, …
  – (plus non-family relationships)

• Tools getting better
Sources

• Researchers believe things with reason
  – Sources, information, evidence, weight, arbitration, inference, ...

• Tool support here still limited
  – Often free–form text
  – Few (e.g., Evidentia) give more structure
Attachments

• Not all “sources” are the source of some belief
  – Photographs, anecdotes, recordings, correspondence, reminiscences, …

• Rapid increase in tool support recently
Stories

• The real truth fit into a narrative
• Many reasons to fit a narrative to the reconstructed past too

• Limited tool integration
  – Trend: story = attachment
  – Few (e.g., Stemma) give more connection to data
Standard Model

- Data structured like ground truth
  - A tree
  - A person relationship graph
- Everything else hangs off that core
  - Sources, attachments, stories, etc
Standard model will fail

- Collaboration
- Split and Merge
- Uncertainty
- Poor memory
- Story Impact
- Cross–tree trends
- Machine learning
Artifacts are things other users could locate or verify (archival). Testimony is from someone other than the user but not archival nor verifiable. User is for things that originate with the user in question directly.

Claims whose sources are not inferences should only reference claims with the same source. Things might include people, places, events, other documents, or anything else a source refers to. Properties of things, including their type, are separate because in general they might be wrong.

Inferences and conflicts might use single-use special-case rules. An inference should only be the source for claims in its consequent. Most conflicts happen between the properties and connections that are inherited by matched nodes.

Decision nodes (belief, refutation, and reliability) are still being designed. This presentation represents just one possible design.

A belief represents what a user sees as "their tree;" it might also be visible to, but not editable by, other users.
Source vs Claim

• Source:
  – Where an idea came from
  – E.g., a document, conversation, personal belief, logical inference, ...

• Claim:
  – The idea that came from it
  – E.g., these two people are brothers, this event happened on this date, ...
Kinds of Claims

• Claim: a person existed
  – (or an event, or a place, …)
  – Many other claims in terms of that

• Claim: these things are related
  – Brothers, happened—at, before, participated—in, …

• Claim: this thing has this property
Matches

• Assertions that a set of claims are about the same thing
  – “The Henry in this document is the same as the Henry in that one”
  – Can be in–document

• (see LifeLines, DeadEnds, etc)
Interferences

• Inferences are an important source
  – Research = search + inference

• Rule + application
  – Not always able to articulate rule
  – Can usually articulate antecedents

• Everything can be sourced
Conflict

• Conflicting ideas are natural
  – Even logical impossibilities, like \( A=B\neq C=A \)
  – Conflicting belief \( \neq \) invalid data

• Conflict resolution = inference
  – Rule: logical inconsistencies aren’t true
Belief, Mutability, Sharing

- Belief = set of other nodes
- My belief ≠ your belief

- All nodes immutable
  - Change = make new, adjust belief

- Collaboration = sharing nodes
Artifacts are things other users could locate or verify (archival). Testimony is from someone other than the user but not archival nor verifiable. User is for things that originate with the user in question directly.

Claims whose sources are not inferences should only reference claims with the same source. Things might include people, places, events, other documents, or anything else a source refers to. Properties of things, including their type, are separate because in general they might be wrong.

---

**Source**
- provenance: Source?
- content: Datum

**Testimony**
- testifier: OutRef

**Artifact**
- citation: OutRef

---

**Inference**
- rule: InferenceRule
- antecedents: Claim[]

**InferenceRule**
- infer: Claim[] Claim[]

---

**Pair**
- y: Claim
- x: Claim
- xIsToY: String

**Group**
- group: String
- members: Claim[2+]

---

**Match T**
- allSame: T[2+]

---

**Belief**
- believed: Node[]
- owner: User
- asOf: Date
- previous: Belief

**Refutation**
- false: Node
- because: Conflict | String

---

**ReliabilityComparator**
- compare: Conflict x Conflict x Context
- boolean

---

**Infer and Infer**

**Detect Conflicts**

**Locate and Parse**

**Incorporate**

---

Inferences and conflicts might use single-use special-case rules. An inference should only be the source for claims in its consequent. Most conflicts happen between the properties and connections that are inherited by matched nodes.

Note
- about: Node
- content: Datum
- creator: User

Decision nodes (belief, refutation, and reliability) are still being designed. This presentation represents just one possible design.

A belief represents what a user sees as "their tree;" it might also be visible to, but not editable by, other users.
Pros of this model

- Collaboration (princess, Henry)
- Split and Merge
- Uncertainty
- Poor memory
- Story Impact
- Cross–tree trends
- Machine learning
Difficulties

• Existing data lacks information needed to change to this data model
  – Change logs come close…
• Some parts of model open to debate
• Much can be automated in theory
  … but how much work is it?
• Change always brings resistance
Questions?

(see http://fhiso.org/call-for-paper-submissions
CFPS 4 and its descendants for more details)