System For Identity Linking and Research Collaboration

Bill Harten

February 28, 2002

Personal Ancestral File, Ancestral File, and the industry that followed took a shortcut, choosing to deal with only tidy, final conclusions. They sidestepped the harder job of storing, linking, and sharing evidence, and of representing conflicting and changing opinions in a community of sovereign researchers. A conclusion-only database offers no basis for future evidence evaluation: no researcher can build on the work published by another.

Databases which support only a single-valued, ideal view require us to make unfounded decisions that *merge identities* arbitrarily without justification. They replace true ambiguity with false simplicity. Yet we require a *best view* with single values to show our opinions at a point in time. How can we faithfully represent evidence and conflicting conclusions in collaboration?

A working system is demonstrated that implements *identity linking*, an alternative to identity merging, including a comprehensive model for sharing genealogical evidence extracts and conclusions-in-progress. This model enables collaboration among researchers who share dynamic conflicting information. It provides three essential views: (1) an evidence identity, (2) combined matching identities with possible conflicts, and (3) single-valued best view with simple conflict resolution. Links can span databases, allowing one researcher to link to another researcher's database dynamically, reflecting updates as it changes. Also, one researcher can register disagreement with another's conclusions without compromising the other's sovereignty.

In this model, data is stored in *identity* records, which represent a person identified in an evidence record, or in *individuals.* An individual represents a distinct real person, in the opinion of a researcher. Both use the same syntax.

Identities and individuals are represented in the form of tagged, structured text, not relationally, although a relational DBMS may be used to store and index them. The syntax is a string of tag and value pairs, rendered left to right, like GEDCOM but without line delimiters or level numbers.

Tags consist of the family links **father**, **mother**, and **child**, the identity link **tie** (tying identities together), **source** links, and basic events **birth**, **marriage**, **death**, and **event**, plus user-definable extensions of the four basic event tags. Users may also define abbreviations or translations of these tags in any language, but these are replaced with standard tags when exported. Event dates and places are parsed contextually, eliminating DATE and PLACE tags. Tags delimit values. Relationship tags delimit individuals and identities in a compound record. This tagged text format has also been used as an efficient data entry format, also suitable for voice data entry, but that is another discussion. Fancy editors can translate this representation to/from other desirable screen representations. Alternative storage representations are also conceivable.

The value of a link tag consists of a global database reference number (which maps to a Internet URI/URL) plus a colon delimiter, and a unique record number in the referenced

database. For example, “3:239” refers to record 239 in database 3. The reference “468” refers to record 468 in the same database as the record containing the reference.

Links represent an insertion of the text string from the referenced record into the text string of the record containing the reference. The substitution occurs dynamically as a view is generated, accessing databases across the Internet. Multiple links and multiple tags with conflicting values are inserted in order of decreasing probability, in the researcher’s opinion, left to right.

Various views may ignore certain links or tags, and may format the information as desired. In views that allow only a single value for an item, such as a *best view*, the view mechanism uses only the first item of a given type found by reading left to right, and ignores subsequent items of the same type.

Consider the following evidence identities in databases 3 and 5:

3:239 identity Tom Jones **birth** 1903 Ohio **marriage** 1922

5:330 identity Thomas Jones **birth** 1901 Ohio **death** 1946

and the conflicting opinions of two researchers creating databases 6 and 7 respectively:

6:101 individual Thomas Jones **birth** 1901 Ohio **tie** 3:239 **tie** 5:330

7:333 individual tie 5:330

7:334 individual tie 3:239

Researcher 6 believes that the two evidence records identify the same person, and also chose to override the name and birth data from record 3:239 with preferred values by placing them to the left of the identity links.

In disagreement, Researcher 7 decided that these are two separate persons and created separate individual records, with one identity link each and no overrides.

To dynamically construct a view of record 6:101, the software first substitutes links with corresponding text from other databases to form

6:101 individual Thomas Jones **birth** 1901 Ohio **tie** 3:239 **identity** Tom Jones **birth** 1903 Ohio **marriage** 1922 **tie** 5:330 **identity** Thomas Jones **birth** 1901 Ohio **death** 1946

This expanded record is a current, comprehensive view of available information on this individual in Researcher 6’s opinion, including conflicts, and is the essential view used in evaluating new evidence.

Building on the expanded record, a family group sheet or pedigree chart *best view*, which allows only single values for birth, death and such, would read left to right and stop looking for birth upon finding **birth 1903 Ohio** from the individual record, etc.

Yet another Researcher 8 might decide to rely on Researcher 6’s work for one branch of the family tree, except for a certain death date in this example, by creating something like the following:

8:932 individual Mary Jones **father** 933

8:933 individual Thomas Jones death 06 JUN 1946 **tie** 6:101 child 932

A pedigree view beginning with Mary Jones in record 8:932 will automatically show the very latest changes for Thomas made by researcher 6, even looking further back in pedigree. This eliminates the miserable headache of reconciling with your cousin's updated GEDCOM files, but brings the occasional headache of the remote database going down. Repositories could help minimize that problem.

Requirements for this concept were articulated in 1995. Development commenced privately in 1997, and was suspended in 1999 until more funding becomes available.