Linking Families with Enriched Ontologies

David W. Embley[1], Stephen W. Liddle[2],
Deryle W. Lonsdale[2], and Scott N. Woodfield[1,2]

[1] FamilySearch International, Lehi UT 84043, USA
[2] Brigham Young University, Provo UT 84602, USA

1 Introduction

Using enriched ontologies, we address a well known and challenging problem: record linkage of historical records for inter-generational family-tree construction. An enriched ontology enables extraction of birth, death, and marriage records via linguistic grounding, curation of record-comprising information with pragmatic constraints and cultural normatives, and record linkage by evidential reasoning. The result is a fully automatic reconstruction of family trees from book-length historical documents.

2 Ontological Enrichments

Linguistic Grounding. A user programs an extraction engine, GreenQQ [5], by giving examples. GreenQQ generates templates from given examples to classify entities in a book's text stream. A post-processing program groups classified entities into relationships and populates the ontology's conceptual model.
Pragmatic Constraints. Pragmatic constraints facilitate a semantic analysis of GreenQQ's syntactically extracted information [11]. For family reconstitution, if a potential merge of two records violates a pragmatic constraint (e.g. asserts that a child was born to a mother who was deceased), the merge is rejected.
Cultural Normatives. Information obtained by "reading between the lines" [4] is invaluable in family reconstitution. Using cultural normatives, missing surnames of children can be inferred as can missing female spouse names. Along with life realities, cultural normatives also aid in estimating missing birth dates (e.g. by knowing a christening date or a marriage date or a child's birth date).
Evidential Reasoning. Family tree construction consists of identifying individuals and establishing spouse and parent-child relationships. Extracted records comprise this information but for each individual i the records that pertain to i must be identified and merged. Identifying which persona records to merge is a record linkage problem whose resolution is aided by evidential reasoning [3].

3 Family Linkage

Automated record linkage has been studied for more than 60 years [9] and continues to be studied with varying degrees of success [1, 2, 6]. Standard approaches consist of three phases: input preparation, blocking, and within-block matching. Ontology enrichments provide the basis for enhancing each phase of record linking: input preparation is more extensive, blocking is governed by shallow matching based largely on inferred evidence, and final matching (including cross-block matching) is deep—based on an extensive use of garnered ontological knowledge.

We conducted field experiments on three books: Ely [10], Kilbarchan [7], and Miller [8]. For each, we ran the full automation pipeline from GreenQQ extraction through deep-match record equivalence-class construction. We then merged the deep-matched records in the same equivalence class, linked them inter-generationally, and produced GedcomX files representing family trees. Table 1 gives statistics for shallow-match blocking, and Table 2 gives statistics for deep matching. By sampling and checking results for Ely and Miller, we estimated the percent of false positives (erroneous deep-match equivalence classes) and false negatives (deep-match equivalence classes with missing records). Because of the nature of the Kilbarchan book, only a few deep-match equivalence classes were produced, allowing an exhaustive check. Table 3 shows the results.

Table 1. Generation of Shallow Match Equivalence Classes (Blocking).

	# records	**execution**	**surn. inferd**		**birth dates**		**# eq. cls. size**		
Book (pages)	**found**	**time (ms)**	**birth**	**mar.**	**extr.**	**est.**	**1**	**2–9**	**10^+**
Ely (432–700)	8,976	16,228	2,731	3,038	4,427	3,895	5,415	1,208	8
Miller (7–395)	11,439	30,037	1,532	2,573	2,818	8,303	7,749	1,554	1
Kilb. (4–127)	8,814	11,087	4,043	2,064	1,103	6,224	5,049	1,174	15

Table 2. Generation of Deep Match Equivalence Classes.

	execution	**# of size**			**# recs.**	**# recs. pushed across blocks**		
Book (pages)	**time (ms)**	**1**	**2–9**	**10^+**	**rejctd**	**unmrgabl**	**rejctd**	**not confidnt**
Ely (432–700)	145,095*	6,479	865	2	146	3,312	1	3,615
Miller (7-395)	120,138	10,164	572	41	0	2,092	5	6,493
Kilb. (4–127)	97,520	8,334	12	0	438	7,819	0	10,955

* Without blocking, an estimated 5 days would have been required to process Ely (432-700).

Table 3. Deep Match Equivalence Class Accuracy.

	false	**false**	**# checked**	**Accuracy**		
	positives	**negatives**	**(Accuracy)**	**Recall**	**Precision**	**F-score**
Ely (432–700)	2	16	80	83%	98%	90%
Miller (7-395)	9	4	80	95%	89%	92%
Kilb. (4–127)	12	0	8,346	100%	99.86%	99.93%

Deep-match equivalence class F-scores for [7], [8], and [10] ranged from 90% to 99%. By merging the records in each equivalence class and reconstituting parent-child links across merged records, we are able to automatically create inter-generational family trees for these books with an accuracy in the 90th percentile.

References

1. Abramitzky, R., Mill, R., Perez, S.: Linking individuals across historical sources: a fully automated approach (2018), working paper #1031
2. Bailey, M., Cole, C., Henderson, M., Massey, C.: How well do automated linking methods perform?—lessons from U.S. historical data (2019), working paper
3. Embley, D., Liddle, S., Lonsdale, D., Woodfield, S.: Inter-generational family reconstitution with enriched ontologies (2020), available at https://deg.byu.edu/
4. Embley, D., Liddle, S., Park, J.: Increasing the quality of extracted information by reading between the lines. In: Comyn-Wattiau, I., du Mouza, C., Prat, N. (eds.) Ingénierie et management des systèmes d'information—Mélanges en l'honneur de Jacky Akoka (December 2016)
5. Embley, D., Nagy, G.: Green interaction for extracting family information from OCR'd books. In: Proceedings of the 13th IAPR International Workshop on Document Analysis Systems. (DAS 2018), Vienna, Austria (March 2018)
6. Feigenbaum, J.: A machine learning approach to census record linking (2016), available at http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaumcensuslink
7. Grant, F.: Index to The Register of Marriages and Baptisms in the PARISH OF KILBARCHAN, 1649–1772. J. Skinner & Company, LTD, Edinburgh, Scotland (1912)
8. Miller Funeral Home Records, 1917 – 1950, Greenville, Ohio. Darke County Ohio Genealogical Society, Greenville, Ohio (1990)
9. Newcombe, H., Kennedy, J., Axford, S., James, A.: Automatic linkage of vital records. Science **130**, 954–959 (October 1959)
10. Vanderpoel, G.: The Ely Ancestry: Lineage of RICHARD ELY of Plymouth, England. The Calumet Press, New York, New York (1902)
11. Woodfield, S., Lonsdale, D., Liddle, S., Kim, T., Embley, D., Almquist, C.: Pragmatic quality assessment for automatically extracted data. In: Proceedings of ER 2016. pp. 212–220. Gifu, Japan (November 2016)