Inter-Generational Family Reconstitution with Enriched Ontologies ??

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

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

Abstract. What's the problem?: reconstitution of family relationships within a community of interest. Why's the problem a problem?: the desire to prosopographically analyze extended family relationships within a community of interest. What's the solution?: automated family reconstruction from historical records using conceptual models enriched ontologically with linguistic grounding, pragmatic constraints, cultural normatives, and facilities for evidential reasoning. Why's the solution a solution?: (1) a fully automatic reconstruction (from fact extraction through family tree construction) of The Ely Ancestry (12,710 person-mention instances and 29,732 items of information for these persons); a sampling of the reconstituted family tree checked against the ground truth as given in [10]: ??%recall and ??%precision; (2) a fully automatic reconstruction of families in the Kilbarchan Parish Community, 1649–1772, from marriage and christening/birth events recorded independently by parish priests [8]; although no ground truth for this reconstruction exists, a sampling of 100 inter-generational relationships from different family lines all appeared to be correct with respect to biological and cultural norms; (3) similar to (2) for Miller Funeral Home Records [9]. (Miller is more likely to be successful than Kilbarchan because each person typically has more information on which to judge duplicates.)

Keywords: Prosopography · Enriched ontologies · Linguistic grounding · Pragmatic constraints · Cultural normatives · Evidential reasoning.

1 Introduction

... motivating examples (Ely 575 Rev. Ben family + Harriet family, Miller ESTHER, Kilbarchan with run-on family) ... list and illustrate our ontological enrichments

... contribution: enriched ontologies provide a sound conceptual-modeling basis for inter-generational family reconstitution

2 Ontological Enrichments

2.1 Linguistic Grounding

... ontological document reading [3]

... entity extraction via data frames, but with GreenQQ [5] entity-extraction rules and style families for dates and names (and possibly places); syntactic patterns; one-shot learning (i.e. only need one training example to create a rule; for semi-structured family history books, a couple of examples for each entity is sufficient).

... relationship extraction via GreenQQ grouping wrt record views of a target ontology ... extraction ontology diagram ... Record Normalization, a principled approach that generalizes: (Person, Name, GenderDesignator, BirthDate, BirthPlace, ChristeningDate, ChristeningPlace, DeathDate, DeathPlace, BurialPlace), (Person, (Spouse, MarriageDate, MarriagePlace)*); (Person1, Person2, (Child)*)—illustrate with Ely. The main property of a record is that every property has a direct relationship to the object the record represents: Individual works for Miller—and perhaps we should lead with Miller because it illustrates the main point about records, namely that each item of interest for forming directly pertains to the object a record represents (formalize this notion of a *record* with a definition). Cardinality constraints guide the process of converting head-to-head groupings into records (e.g. missing child head in Ely children yields two BirthDates)

... target ontology population via record-view integration ... wrt personas ... we use *persona* to denote mentions of a person in our books, usually there are several personas based on the roles they play—parent-of, child-of, spouse-of, person-of-focus.

2.2 Pragmatic Constraints

... semantic constraint reconciliation ... bad relationships fixed ... divisions: (1) modification (e.g. sorting out children for Rev. Ben's spouses); (2) retraction (e.g. too many parents in Kilbarchan); (3) expansion (see subsection 2.3)

... report violations if can't fix and do so with precision ... human intervention (if desired) with mark-up of violations of semantic constraints, authority checks of names and places, violations of biological and cultural norms ... cross-check of source and extracted/inferred info with COMET ... probabilistic, the why's explained, direction for resolution

... for more details, see pragmatic quality assessment for automatically extracted data [11] and ontological deep data cleaning [12]

2.3 Cultural Normatives

... standardization of names, dates, and places ... authority checks ... fix OCR errors (e.g. names and places based on name and place authorities; dates based on assumptions about standard forms of dates—actually done earlier so that dates can be parsed)

... inference: gender, surnames of children and spouses based on birth and marriage, inferred inverse spousal and parent-child relationships (not stated but inferred by reading between the lines [4])

... coding needed: none

2.4 Evidential Reasoning

... record linkage references: [7], [1], [2]

... approach: input preparation (the above + producing canonical values and birth date estimates for those without birth dates + information content sorting), shallow equivalence class creation (blocking), deep equivalence class creation (match equivalance class with primary persona, Tree-Sweeper red and yellow flags [12]), merge to form persons as nodes in a family tree.

Input Preparation As described in Sections 2.1–2.3 we have created for each *persona*—each mention instance of a person in a document—a description consisting of extracted and standardized date and place facts for birth, marriage, and death events and all extracted and inferred "one-hop" family relationships to parents, spouses, and children. Conceptually, we list the personas ordered by how information rich their persona description is (ordered most to least). For purposes of semantically judging persona matches, we add semantically curated information to persona descriptions: for Dates: a julian date range with precision dependent on the extracted date (e.g. for "January 1853" the julian date range is "1853001-1853031"); for Places: longitude and latitude values (not currently implemented); for Names: labeled name parts: title(s), first name(s), last name(s), and suffix(es) such as "Jr." and "Sr.". Because of their importance in matching, we also add estimated birth dates for every persona for whom no birth date has been extracted. We estimate based (1) on any extracted birth, death, and marriage event dates (e.g. an estimated birth date being normally a few weeks before or even right up to the date of christening) or (2) on extracted birth dates of one-hop relationships (e.g. a first child being born 20 years or so after a mother's birth).

Shallow-Match Equivalence Class Construction The equivalence-class relationship for shallow-match equivalence classes is "is a plausible match with the first persona placed in the equivalence class." Selecting from the information-rich-ordered list of persona descriptions, we form an ordered list of equivalance classes with each equivalence also being ordered—all ordered most to least in information richness. In greedy fashion, we add each persona description to the first equivalence class to which it has a plausible match with the first persona description. Our criteria for being a plausable match is (1) that standardized birth names (extracted or inferred) are identical and (2) that birth julian date ranges (extracted or estimated) overlap. Blocking techniques normally require that all potential matches appear in the same block. Our blocking does not because any persona that does not deep-match (as described next) with the first persona in a shallow-match equivalence class is pushed downstream in the ordered shallow-match equivalence-class list. What is required, however, is that no persona is pushed downstream passed the equivalence class q in which the persona is a match with the first persona in q. In our application, (1) we can count on document authors being consistent in the way they render person names (thus,

although not generalizable, "identical" works for our application), and (2) we can reject matches when birth dates are not reasonably close (even as measured by our generous range spans when they must be estimated).

Deep-Match Equivalence Class Construction The equivalence-class relationship is "is a match." The check is deep and based on the idea that if merged, then it makes sense semantically and is near certain probabilistically. The approach is to check, in order, whether a persona matches with each of the personas that preceed it in the ordered list in a shallow-match equivalence class. Match criteria includes all the semantic-reasoning checks described in Section 2.2 and also includes semantic-compatibility checks of all one-hop relationships to personas in both persona descriptions if they were to be merged. Examples:

1. If persona P_1 has mother M_1 and persona P_2 has mother M_2, then M_1 must shallow-match with M_2.
2. If persona P_1 has spouse S_1 and persona P_2 has spouse S_2 and S_1 does not shallow-match with S_2, then the marriage date ranges of of spouse S_1 and S_2 must not overlap.
3. If persona P_1 has child C_1 with spouse S_1 and persona P_2 has child C_2 with spouse S_2 and S_1 shallow-matches S_2 and C_1 does not shallow-match with C_2, then the birth dates of C_1 and C_2 must be more than nine months apart.

Observe that all these statements are implications as are the reasoning rules described in Section 2.2. They are also all based on probability distributions derivable from information repositories like those in the Family Tree hosted online by FamilySearch International [6]. ... when the system throws a red flag for any rule, we reject the match ... yellow-flag warnings are accumulated and analyzed along with green-flag confirmations to determine whether the personas match as follows ...

Inter-Generational Family Tree Generation ... merge ... recursive untenable resolution and our approach to inter-generational tree reconstitution ... approach description needed (SNW)

3 Field Experiments

... assessment of results (SNW, DWE, DWL, SWL)

3.1 The Ely Ancestry

... Ely book: automatic Ely Family Tree reconstitution with spot check against the ground truth in the book

3.2 Kilbarchan Parish Community

... Kilbarchan: automatic reconstruction of the Kilbarchan community as represented in the Kilbarchan Parish Register of Marriages and Baptisms 1649-1772

3.3 Miller Funeral Home Records

4 Concluding Remarks

...

Future Work: ... fine tuning of distributions ... generalizing for when book-assumptions do not hold

References

1. Abramitzky, R., Mill, R., Perez, S.: Linking individuals across historical sources: a fully automated approach (2018), working Paper No. 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.: Ontological document reading: An experience report. Enterprise Modelling and Information Systems Architectures: International Journal of Conceptual Modeling pp. 133–181 (February 2018)
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. FamilySearch. http://familysearch.org
7. Feigenbaum, J.: A machine learning approach to census record linking (2016), available at http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaumcensuslink
8. Grant, F.: Index to The Register of Marriages and Baptisms in the PARISH OF KILBARCHAN, 1649–1772. J. Skinner & Company, LTD, Edinburgh, Scotland (1912)
9. Miller Funeral Home Records, 1917 – 1950, Greenville, Ohio. Darke County Ohio Genealogical Society, Greenville, Ohio (1990)
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)
12. Woodfield, S., Seeger, S., Litster, S., Liddle, S., Grace, B., Embley, D.: Ontological deep data cleaning. In: Proceedings of the 37th International Conference on Conceptual Modeling. (ER 2018), Xi'an, China (October 2018)