

Inter-Generational Family Reconstitution with Enriched Ontologies

David W. Embley^{1,2}, Stephen W. Liddle¹,
Deryle W. Lonsdale¹, and Scott N. Woodfield¹

¹ Brigham Young University, Provo UT 84602, USA

² FamilySearch International, Lehi UT 84043, USA

Abstract. Enriching ontologies can measurably enhance research in digital humanities. Support for this claim is shown by using an enriched ontology to attack a well known and challenging problem: record linkage of historical records for inter-generational family reconstitution. 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. Using three historical record books containing a total of 29,229 extracted records, the enriched ontology links records with high accuracy: F-scores in the 90% range for all three books.

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

1 Introduction

Genealogical family relationships often form the basis for prosopographical research within a community of interest. Genealogical research focuses on internal family relationships, whereas prosopographical studies focus on external relationships of family members to community services, employment networks, marriages, and social and religious groups. For historical group studies and other applications such as inherited-disease research and assisting genealogy enthusiasts, we show in this paper that augmenting ontologies with rich real-world constraints and cultural normatives can lead to a fully automatic reconstitution of intergenerational family-lineage trees from information automatically extracted from semi-structured records found in community-oriented family history books.

Figures 1, 2, and 3 are text snippets from three family history books. Each mention of a person in a document is a *persona*. The objective is to discover the intergenerational relationships given the *persona records*—the personas and their related information. This requires (1) discovery of parent-child relationships among the personas and (2) discovery of which personas refer to the same person—a persona record linkage problem. Examples:

- In Figure 1 the persona “Rev. Ezra Stiles Ely” matches the persona “Ezra Stiles Ely” despite their having different spouses and children born 20 years

apart. Elsewhere in *The Ely Ancestry* [10] is a persona with birth, death, marriage, and parent-child information fully consistent with Ezra’s having the two mentioned wives along with their marriage and death dates that form time windows in which the two children Ben and Harriet were born.

- In Figure 2 persona “TEEGARDEN, WM. WALTER” matches persona “W.W. TEEGARDEN” despite the name variations.
- In Figure 3, is the persona “John Adam” who was christened on 30 May 1652 the same as persona “Adam, John” married to Jean Reid? His age when Jean’s son John was born would have been about 21—a likely age for a father of a first child. Or, is he the same person as the John Adam who was married to Agnes Andro? A marriage in 1679 would have been when he was about 27—not unreasonable. Or, is he neither of these two?

243327. Rev. Ben Ezra Stiles Ely, Ottumwa, Ia., b. 1828, son of Rev. Ezra Stiles Ely and Mary Ann Carswell; m. 1848, Elizabeth Eudora McElroy, West Ely, Mo., who was b. 1829, d. 1871, dau. of Abraham McElroy and Mary Ford Radford; m. 2nd, 1873, Abbie Amelia Moore, Harrison, Ill., who was b. 1852, dau. of Porter Moore and Harriet Leonard. Their children :

1. Elizabeth B., b. 1849.
2. Ben-Ezra Stiles, b. 1856.
3. George Everly Montgomery, b. 1858, d. 1877.
4. Laura Elizabeth, b. 1859.
5. LaRose DeForest, b. 1861.
6. Charles Wadsworth, b. 1863.
7. Mary Anita, b. 1865.
8. Francis Argyle, b. 1876.

243320. Harriet Clarissima Ely, b. 1848, dau. of Ezra Stiles Ely and Caroline Thompson Holmes; m. Beale Steenberger Blackford, Parkersburg, W. Va. Their children:

1. Caroline Holmes Ely.

Fig. 1. Text Snippet from *The Ely Ancestry* [10]—family expansion and migration beginning in Boston, Massachusetts, USA (~1650–1900).

We call our ontology-enriched record linking system *OntoLink*. Its putative contributions include:

1. ontology enrichments: linguistic grounding (Section 2.1), pragmatic constraints (Section 2.2), cultural normatives (Section 2.3), and evidential reasoning (Section 2.4); and
2. (a) a shallow-match blocking technique that remains efficient but allows for cross-block matches and (b) a deep-match, evidential-reasoning technique that not only successfully matches personas but also yields the reasoning behind matches, mismatches, and low-confidence possible matches (Section 3).

2 Ontological Enrichments

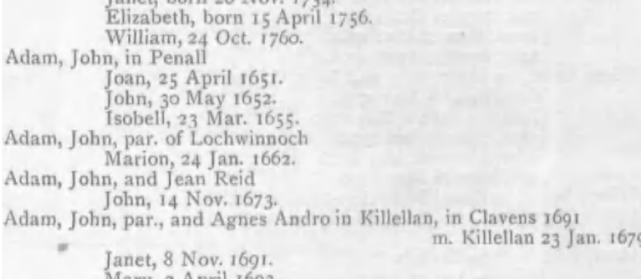
The OntoLink pipeline for automatically constructing inter-generational family lineage trees depends critically on the ontological enrichments we propose. It

TEEGARDEN, CATHERINE 404 West Fourth St d 6 May 1941 1:00p.m. Wayne Hosp
 Greenville OH BD Greenville 8 May 1941 b 20 Nov 1865 Greenville Twp Dke
 Co OH age 75-5-16 f JOHN SWAC? HERSHEY Lancaster Co PA m ANNA YOUNG
 Lancaster Co PA widowed housewife sp W.W. TEEGARDEN physician Dr Gil-
 bert Sayle religion Evangelical & Reform funeral 8 May 1941 2:30p.m.
 Thursday Evangelical & Reform Rev E.V. Louks survived by 3 sons ROLAND
 Sidney, HAROLD Washington NC, and CHESTER Albany NY, 1 daughter LORENE
 TEEGARDEN Cincinnati

TEEGARDEN, LORENE d 2 Nov 1946 Washington D.C. residence for 2 years BD Green-
 ville Cem Greenville OH 5 Nov 1946 single 3 brothers HAROLD of Washing-
 ton D.C., ROLAND of Sidney OH, CHESTER of NY

TEEGARDEN, WM. WALTER d 6 July 1936 Greenville BD Greenville 8 July 1936
 b 17 July 1862 Brown Twp Dke Co OH age 74-11-19 f MOSES TEEGARDEN Dke Co
 OH m HANNAH DAY MENDENHALL sp CATHERINE TEEGARDEN

Fig. 2. Three *Miller Funeral Home Records* [8]—patrons of a funeral home in Greenville, Ohio, USA (~1910–1950).



Elizabeth, born 15 April 1756.
 William, 24 Oct. 1760.
 Adam, John, in Penall
 Joan, 25 April 1651.
 John, 30 May 1652.
 Isobell, 23 Mar. 1655.
 Adam, John, par. of Lochwinnoch
 Marion, 24 Jan. 1662.
 Adam, John, and Jean Reid
 John, 14 Nov. 1673.
 Adam, John, par., and Agnes Andro in Killellan, in Clavens 1691
 m. Killellan 23 Jan. 1679
 Janet, 8 Nov. 1691.
 Mary, 8 April 1668.

Fig. 3. *Parish of Kilbarchan* Text Snippet [6]—a community of Scottish worshippers (~1640–1780).

begins with information extraction based on linguistic grounding and proceeds through information curation based on the semantics of pragmatic constraints and cultural normatives which prepares the extracted information for family tree construction via evidential reasoning.

2.1 Linguistic Grounding

A user programs an extraction engine, GreenQQ [4], by giving examples. GreenQQ generates templates from given examples to classify entities in a book’s text stream. Then, with respect to a chosen “head” class, GreenQQ groups identified entities into records from which OntoLink can generate object and relationship instances that populate the conceptual model underlying an ontology.

For example, from Figure 3 a user may give GreenQQ the sample text “, 30 May 1652.\n”. With the date marked as belonging to the class *ChristeningDate*, GreenQQ generates the template pattern “, ChristeningDate:[NUM1or2 CAP NUM4] . EOL”. When applied over the entire book, all text snippets that satisfy this pattern are classified as christening dates. Similarly templates can be created for birth dates, which, as seen in Figure 3, follow the literal “born” and also for the names of the children with birth and christening dates, which have the

template pattern “SOL Name:[CAP] , (NUM1or2 | born)”. Then, with “Name” chosen as the “head” class, GreenQQ can group classified entities into christening and birth records—hundreds of them in the Kilbarchan book.

2.2 Pragmatic Constraints

Pragmatic constraints facilitate a semantic analysis of GreenQQ’s syntactically extracted information. In Figure 1, for example, it is impossible to syntactically associate Rev. Ely’s children with their proper mother. Pragmatically, however, Francis cannot be a child of Elizabeth since she was dead when Francis was born. OntoLink identifies and, when possible, rectifies these kinds of errors [12].

For inter-generational family reconstitution, if a potential merge of two personas violates a pragmatic constraint, OntoLink raises a red flag and rejects the merge. If, for example, the parents of two potentially merged personas do not properly correspond, a merge would violate the constraint that a person can have only two parents. In addition to signaling impossibility, violations can also flag improbability. In Figure 3, for example, a merge of John Adam christened on 30 May 1652 and John Adam of Lochwinnoch would mean that the child Marion, christened on 24 Jan. 1662, would have been born when John was only about 10 years old—improbable.

OntoLink also raises red flags when corresponding values such as names and birth dates in potentially matching personas are not close enough to be considered equivalent. In Figure 1, for example, based on the name, Mary Ann Carswell is clearly not the same person as Caroline Thompson Holmes even though they have the same spouse. On the other hand, W.W. TEEGARDEN in Figure 2 is the same person as TEEGARDEN, WM. WALTER even though their name parts have different spellings and orderings.

2.3 Cultural Normatives

Information obtained by “reading between the lines” [3] is invaluable in inter-generational family reconstitution. Surnames of the children in Figures 1 and 3 can be ascertained knowing cultural normatives. Likewise, in Figure 2 it is clear that TEEGARDEN is CATHERINE’s married surname and that her maiden surname is HERSHEY. Reading between the lines, it is also possible to garner missing information. Cultural normatives and local religious practice strongly indicate that John Adam’s missing date of birth in Figure 3 was a few weeks prior to his christening.

To make these cultural normatives easy to work with, OntoLink standardizes the text and canonicalizes the text values. It recasts all dates in the form (day, spelled-out month, year), and it canonicalizes them as Julian date ranges so that, for example, Mary Anita’s birth date in Figure 1 is canonicalized as 1865001-1865365. For locations, OntoLink standardizes by ordering administrative levels as usual and canonicalizes by specifying longitude and latitude (not yet implemented, and not used in our field experiments). For names, OntoLink standardizes by giving birth names in their usual capitalization and order, and

canonicalizes by labeling name parts and including titles, married names, and suffixes. For example, TEEGARDEN, CATHERINE in Figure 2 is standardized as “Catherine Hershey” and canonicalized as “Title(s): FirstName(s): Catherine BirthSurname: Hershey MarriedSurname(s): Teegarden Suffix(es):”. If she were a physician, she might have the suffix “M.D.”.

2.4 Evidential Reasoning

Inter-generational family tree construction consists of identifying individuals and establishing spouse and parent-child relationships. Persona records comprise this information but for each individual i the persona records that pertain to i must be identified and merged. Identifying which persona records to merge is a record linkage problem whose resolution requires evidential reasoning.

Automated record linkage has been studied for more than 60 years [9] and continues to be studied with varying degrees of success [1, 2, 5]. Standard approaches consist of three phases: input preparation, blocking, and within-block matching. OntoLink’s 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 is deep—based on an extensive use of garnered ontological knowledge.

Input Preparation. As described in Sections 2.1–2.3, OntoLink creates for each *persona* a *persona record* consisting of extracted and standardized name, date, and location facts for birth, marriage, and death events and all extracted and inferred “one-hop” family relationships to parents, spouses, and children. Recall, in particular, that for every lexical value there is both a standardized value (to aid identity matching) and a canonical value (to aid in measured closeness matching). Because of their importance in matching, OntoLink also adds an estimated birth date, when possible, for every persona for whom no birth date has been extracted. The estimate is based on (1) any extracted christening, death, burial, 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) extracted birth dates of one-hop relationships (e.g. a first child being born 20 years or so after a mother’s birth). The estimated birth date is then set, marked as approximate, and given a date range. For John Adam in Figure 3 christened on 30 May 1652, for example, the computed pair is: (estimated birth date: ~1652109, estimated birth date span: 1652067–1652151). If several estimates are possible, the most precise estimated is chosen.

Blocking (Shallow-Match Equivalence Class Construction). OntoLink orders the persona records by description information richness (most to least). Selecting from this list, it forms an ordered list of equivalence classes with each equivalence class also being ordered by persona-record richness. The equivalence-class relationship is: “is a plausible match with the first persona placed in the equivalence class.” Thus, in greedy fashion typical of standard blocking, we add

each persona record to the first equivalence class to which it has a plausible match with the first persona record. The criteria for being a plausible match are (1) that birth surnames (if any) match within a specified edit distance; that if no surnames, then at least one (if any) of the married surnames match within a specified edit distance; and that the first names weakly correspond (one name sequence subsumes the other, where matching names have the same initial, have an identical initial/first-letter, or have spelled-out or abbreviated names that match within a specified edit distance) and (2) that birth Julian date ranges (extracted or estimated) overlap, or the minimum of the earlier date range to the maximum of the later date range is within five years. Standard blocking techniques normally require that all potential matches appear in the same block. OntoLink’s blocking does not!—because any persona that does not deep match (as described next) with all the preceding personas in the equivalence class are pushed downstream in the ordered equivalence-class list in such a way as to maintain the invariant constraints of the yet unprocessed part of the shallow equivalence-class list.

Matching (Deep-Match Equivalence Class Construction). The equivalence-class relationship is: “is a match.” The check is deep and based on the ideas (1) that if two personas are merged, then the merged persona makes sense semantically and (2) that the evidence for the match is sufficient to yield a high level of certainty. Each persona in a shallow match equivalence class beyond the first is deep-match-checked pairwise against *all* prior personas in the list. This ensures that the match relationship is reflexive, symmetric, and transitive (for those that remain in the equivalence class).

1. A merge of persona P_1 and P_2 is semantically reasonable if (1) neither P_1 nor P_2 individually raises a red flag as explained in Section 2.2, (2) combining corresponding known lexical values into a single value raises no red flag, and (3) a (temporary) merge of P_1 and P_2 raises no red flag. After merging, all non-duplicate person-parent, person-child, and person-spouse relationships of both P_1 and P_2 are added along with one of the two relationships for each duplicate. A relationship is a duplicate if referenced related personas are shallow-equivalent. For example, $Person(P_1)–Spouse(P_3)$ and $Person(P_2)–Spouse(P_4)$ are duplicates if persona P_3 shallow-matches persona P_4 . Red flags may be raised, for example, because there are too many parents, children are born after their mother’s death, or overlapping marriages violate cultural norms, etc.
2. A single red flag rejects a deep match, but the absence of red flags does not confirm a deep match. Constraint-checker-returned probabilities that do not exceed the red-flag threshold can be considered yellow (cautionary) if they tend toward the red-flag threshold and are green (supportive) otherwise. To compute the certainty of a match with no red flags, we invert the probability of a constraint violation for green and yellow flags and determine whether there is enough green-flag evidence to overcome any yellow-flag concerns. OntoLink’s particular technique is a variation of the technique described in

Lawson et al. [7] into which it injects green- and yellow-flag probabilities. A vector $\langle x_1, \dots, x_n \rangle$ is populated with the probability of a match for each lexical attribute and for each name attribute of each person-spouse, person-parent, and person-child relationship (the probability is zero if one or both personas have no value for the attribute). Beforehand, a weight vector $\langle w_1, \dots, w_n \rangle$ is established in which each w_i is the weight the i^{th} lexical comparison should carry for determining a positive persona match. The weights are learned over a ground truth of matched personas.³ The dot product of the probability vector and the weight vector produces a scalar value. Larger values indicate greater confidence in the match. By inspecting proposed matches in our data, we set a threshold that divides the proposed matches into those considered to have a high enough level of certainty to be declared a match and those that do not.

Examples: In the Kilbarchan text snippet in Figure 3 in a run of OntoLink, some of the John Adam personas shallow match, but none deep match. In the Miller text snippet in Figure 2, OntoLink matches the CATHERINE TEEGARDEN personas and matches the two mentions of her husband. The LORENE personas shallow match but (incorrectly) fail to deep match. They match because her brothers align, but OntoLink only considers one-hop relationships and thus misses this vital clue leaving it with insufficient information to be confident of the match. In the Ely text snippet in Figure 1 all three personas with “Ezra Stiles Ely” in their name shallow match, but the constraint insisting that a father and his son not be the same person rejects Rev. Ben Ezra Stiles Ely as being part of the equivalence class of the Ezra Stiles Ely personas.

Inter-Generational Family Tree Generation. OntoLink’s process for establishing persona matches guarantees that persona records in a deep-match equivalence class can be merged. Merged personas contain all the information needed to display an inter-generational family tree as a pedigree chart, a German Ahnentafel, a Chinese Jiapu, or any other desired rendering of a family tree.

3 Field Experiments

We conducted field experiments on three books: Ely [10], Kilbarchan [6], and Miller [8]. For each, we ran the full automation pipeline from GreenQQ extraction through deep-match equivalence-class construction on a server with a 3.00GHz Intel Xeon processor, 32GB of RAM, and 8TB of local storage.

Table 1 gives statistics for generating shallow-match equivalence classes. As explained in Section 2.4, a shallow persona record match loosely compares names

³ In [7], 880 personas of 9,279 were determined to have matches. From this training set, weights were estimated (e.g. 4.6₀₉₀₈ for Birth Year, 4.8₉₄₇₄ for Father’s Surname, 0.0₀₁₇₆ for Birth Town). Lawson et al. argue that these weights should be universal, depending only on the chosen set of attributes. The technique for computing the weights is described by White [11].

and extracted/estimated birth dates. As Table 1 shows, OntoLink’s curation inferred birth surnames for 28.4% of the persona records and inferred married surnames for 26.3%. For persona record birth dates, 28.6% were extracted while 63.0% were estimated with the rest being unknown. Shallow match blocking processed 29,229 persona records and generated 22,173 shallow equivalence classes of various sizes in 57.352 seconds of processing time.

Table 1. Persona Record Shallow Match Equivalence Classes.

| Book (pages) | # persona records | execution time (ms) | surn. infer. | | birth dates | | # eq. cls. size | | |
|----------------|-------------------|---------------------|--------------|-------|-------------|-------|-----------------|-------|-----------------|
| | | | 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. Persona Record Deep Match Equivalence Classes.

| Book (pages) | execution time (ms) | # of size | | | # pers. redfl | # pushed downstream | | |
|----------------|---------------------|-----------|-----|-----------------|---------------|---------------------|----------|--------|
| | | 1 | 2–9 | 10 ⁺ | | unmrgabl | mrgredfl | unconf |
| 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 |

Table 2 gives the statistics for deep matching. From the original 29,229 persona records, OntoLink generated 26,469 deep equivalence classes of various sizes in 362.753 seconds of processing time.⁴ While forming these equivalence classes, OntoLink red-flagged 584 individual personas as being not self-consistent and pushed 34,292 downstream (some multiple times). Of those pushed downstream, 13,223 were unmergable, 6 were red flagged when merged (not self-consistent), and 21,063 were unconfident (lacking sufficient evidence to confidently merge). Since every red-flag error is based on an ontologically specified constraint, a list of red-flag violations constitutes an explanation about why two persona records cannot be merged. When the evidence for a merge is deemed insufficient, a research plan for resolving the merge question can be generated. Yellow/green-flag probabilities associated with pragmatic constraints and cultural normatives and relative attribute weights indicating the importance of each kind of missing information can guide the research plan.

For Ely and Miller, we estimated the percent of false positives (erroneous deep-match equivalence classes) by checking a sampling of them. For a book’s n ordered deep-match equivalence classes with two or more personas, we selected every $(\lfloor n/40 \rfloor)$ th starting with the m th—a randomly chosen number in the range 1–40. If any one of the members of an equivalence did not match all others, the equivalence class was deemed to be a false positive. For Kilbarchan, only 12

⁴ The time savings from shallow-match blocking is $O(n^2)$. We estimate that it would have taken more than 5 days to process *Ely* (pages 432–700) without blocking.

equivalence classes with two or more personas were generated, which we checked exhaustively. Table 3 shows the resulting number of false positives.

Table 3. Persona Deep Match Equivalence Class Accuracy. (**Recall** = $tp/(tp + fn)$, and **Precision** = $tp/(tp + fp)$, where tp is true positives, fp is false positives, and fn is false negatives.)

| | false | false | # checked (Accuracy) | Accuracy | | |
|----------------|-----------|-----------|-------------------------|----------|-----------|---------|
| | positives | negatives | | 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 | 8346 | 100% | 99.86% | 99.93% |

Obtaining the percentage of false negatives (equivalence classes with missing personas) requires a ground truth that is often unreasonably difficult to obtain. However, the Ely book is organized as an inter-general family tree and as such comprises its ground truth, and Miller lists persons alphabetically by surname helping us to know where to look for potential matching persona records. Using the same 40 Ely and Miller equivalence classes selected for checking for false positives plus the first 40 singleton equivalence classes from the ordered list of deep-match equivalence classes, we obtained the results in Table 3, including the Ely and Miller accuracy results. Having previously determined for the Kilbarchan book that no persona records match with sufficient certainty, OntoLink should return every deep equivalence class as a singleton. As Table 2 shows, OntoLink returned 8334 singletons and 12 non-singletons (all false positives), which yields the Kilbarchan accuracy results in Table 3.

Correctness depends on (1) the source documents being error free (having no author-understanding or -recording mistakes, no typing/type-setting mistakes, and no OCR errors) and (2) OntoLink properly capturing and curating document-provided information. In checking a random sample of three pages from each book with a total of 1,022 persona record fields, precision, recall, and F-score for Linguistic Grounding with GreenQQ were respectively 90%, 81% and 86%; and for Pragmatic Constraint identification and rectification were 97%, 87%, and 92%. Checking ten randomly selected persona records from each book with a total of 453 persona record fields that were standardized, inferred, and canonicalized according to Cultural Normatives, the scores were 98%, 98%, and 98%. Observe that the F-scores increase as garnered information is curated along the OntoLink pipeline, indicating the value of ontological semantic enrichment.

4 Concluding Remarks

Deep-match equivalence class F-scores for [6], [8], and [10] ranged from 90% to 99%. Since a collection of all deep-match equivalence classes for a book comprises its family trees, OntoLink was able to automatically create inter-generational family trees for these books with an accuracy in the 90th percentile.

Although much remains to be done—add location information, obtain weights for our application data and determine whether these weights are indeed universal, improve pipeline processing, and do more testing to adjust the set of constraints and fine-tune parameters and thresholds—the results of this preliminary study are promising. Moreover, they support the claim that enriching an ontology with linguistic grounding, pragmatic constraints, cultural normatives, and evidential reasoning can measurably enhance the work of record linkage as a contribution to digital humanities.

Acknowledgements

We are indebted to Emeritus Professor George Nagy, Rensselaer Polytechnic Institute, for the development of GreenQQ.

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., 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)
4. 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)
5. Feigenbaum, J.: A machine learning approach to census record linking (2016), available at <http://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaumcensuslink>
6. Grant, F.: *Index to The Register of Marriages and Baptisms in the PARISH OF KILBARCHAN, 1649–1772*. J. Skinner & Company, LTD, Edinburgh, Scotland (1912)
7. Lawson, J., White, D., Price, B., Yamagata, R.: Probabilistic record linkage for genealogical research. *Brigham Young University Studies* **41**(2), 161–174 (2002)
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. White, D.: A review of the statistics of record linkage for genealogical research. In: *Record Linkage Techniques—1997: Proceedings of an International Workshop and Exposition*. pp. 362–373. Washington DC, USA (1999)
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)