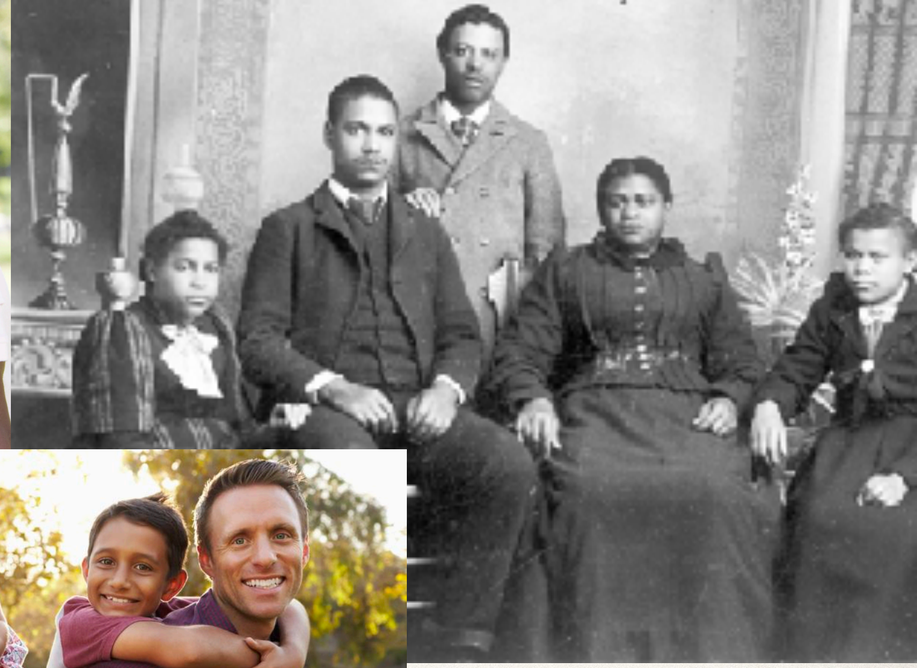
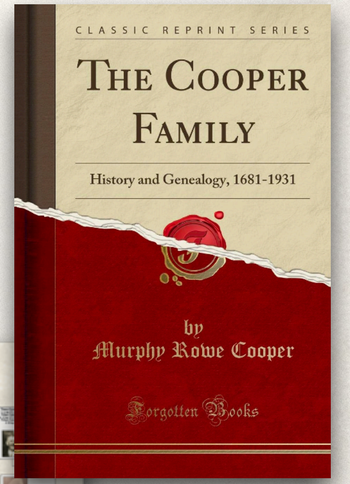




Inter-Generational Family Reconstitution with Enriched Ontologies

David W. Embley, Stephen W. Liddle, Deryle W. Lonsdale, Scott N. Woodfield

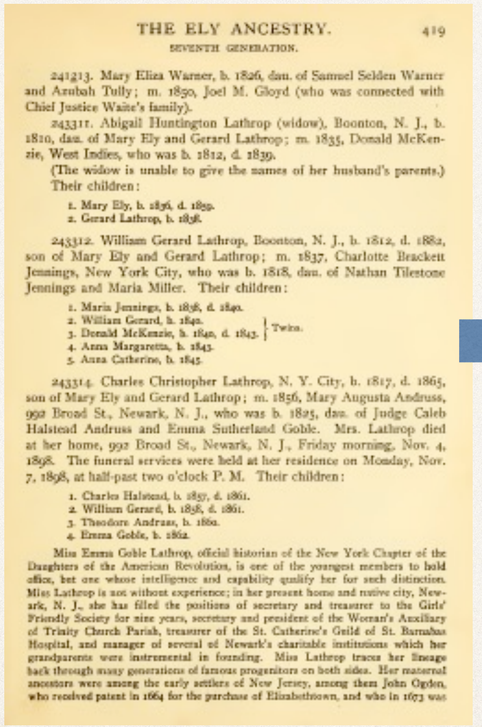
Brigham Young University, Provo, Utah and FamilySearch International, Lehi, Utah



18 October 2021, EmpER 2021 Workshop, St. Johns, Newfoundland, Canada

Historical Record Linkage

- Why study historical family records?
  - To learn about ancestors' history
  - To conduct prosopographical studies
  - To understand one's genetic heritage



THE ELY ANCESTRY. 419

SEVENTH GENERATION.

241313. Mary Eliza Warner, b. 1826, dau. of Samuel Selden Warner and Azubah Tully; m. 1850, Joel M. Gloyd (who was connected with Chief Justice Waite's family).

243311. Abigail Huntington Lathrop (widow), Boonton, N. J., b. 1810, dau. of Mary Ely and Gerard Lathrop; m. 1835, Donald McKenzie, West Indies, who was b. 1812, d. 1839.

(The widow is unable to give the names of her husband's parents.)

Their children:

1. Mary Ely, b. 1836, d. 1859.
2. Gerard Lathrop, b. 1838.

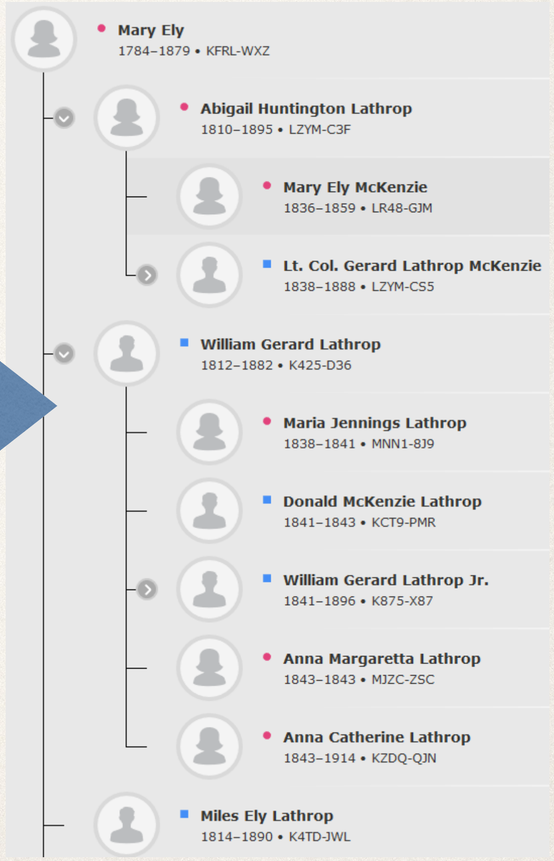
243312. William Gerard Lathrop, Boonton, N. J., b. 1812, d. 1882, son of Mary Ely and Gerard Lathrop; m. 1837, Charlotte Brackett Jennings, New York City, who was b. 1818, dau. of Nathan Tilestone Jennings and Maria Miller. Their children:

1. Maria Jennings, b. 1838, d. 1840.
2. William Gerard, b. 1840.
3. Donald McKenzie, b. 1840, d. 1843. } Twins.
4. Anna Margaretta, b. 1843.
5. Anna Catherine, b. 1845.

243314. Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop; m. 1856, Mary Augusta Andruss, 992 Broad St., Newark, N. J., who was b. 1825, dau. of Judge Caleb Halstead Andruss and Emma Sutherland Goble. Mrs. Lathrop died at her home, 992 Broad St., Newark, N. J., Friday morning, Nov. 4, 1898. The funeral services were held at her residence on Monday, Nov. 7, 1898, at half-past two o'clock P. M. Their children:

1. Charles Halstead, b. 1857, d. 1861.
2. William Gerard, b. 1858, d. 1861.
3. Theodore Andruss, b. 1860.
4. Emma Goble, b. 1862.

Miss Emma Goble Lathrop, official historian of the New York Chapter of the Daughters of the American Revolution, is one of the youngest members to hold office, but one whose intelligence and capability qualify her for such distinction. Miss Lathrop is not without experience; in her present home and native city, Newark, N. J., she has filled the positions of secretary and treasurer to the Girls' Friendly Society for nine years, secretary and president of the Woman's Auxiliary of Trinity Church Parish, treasurer of the St. Catherine's Guild of St. Barnabas Hospital, and manager of several of Newark's charitable institutions which her grandparents were instrumental in founding. Miss Lathrop traces her lineage back through many generations of famous progenitors on both sides. Her maternal ancestors were among the early settlers of New Jersey, among them John Ogden, who received patent in 1664 for the purchase of Elizabethtown, and who in 1673 was



Linking Families

We link families by linking records about the same individual

who was b. 1822, dau. of William Gillingham and Jane McPherson.

24138. Charles Selden, b. 1816, d. 1871, son of Roger Selden and Mary Schuyler Douglas; m. 1845, Hester Gillingham, dau. of William Gillingham and Jane McPherson; m. 2nd, 1862, Elizabeth Keiffer, who was b. 1839.

24331. Mary Ely, Lebanon, Conn., b. 1784, d. 1879, dau. of Zebulon Ely and Sarah Apame Mills; m. 1809, Gerard Lathrop, New York City,

300 THE ELY ANCESTRY.

SIXTH GENERATION.

who was b. 1778, d. 1852, son of Azuriah Lathrop and Abigail Huntington. Their children:

1. Abigail Huntington, b. 1810, m. 1835 Donald McKenzie.
2. William Gerard, b. 1812, d. 1882, m. 1837 Charlotte Brackett Jennings.
3. Mills Ely, b. 1814.
4. Charles Christopher, b. 1817, d. 1865, m. 1856 Mary Augusta Andruss.
5. Ezra Stiles, b. 1818.
6. Mary Cornelia, b. 1820, d. 1879, m. 1840 Jacob Dykeman Vermilye.
7. Charlotte Eliza, b. 1822.

24332. Ezra Stiles Ely, Philadelphia, Pa., b. 1786, d. 1861, son of Zebulon Ely and Sarah Apame Mills; m. 1814, Mary Ann Carswell, who was b. 1792, d. 1842, dau. of Samuel King Carswell and Margaret Means; m. 2nd, 1843, Caroline Thompson Holmes, Virginia, who was b. 1812, dau. of Thompson Holmes and Eliza Allen Stockly. Their children:

1. Margaret Carswell b. 1815, m. 1832 Samuel McClellan.
2. Samuel Carswell, b. 1816, d. 1878.
3. Ezra Stiles, b. 1818.

THE ELY ANCESTRY. 419

SEVENTH GENERATION.

241213. Mary Eliza Warner, b. 1826, dau. of Samuel Selden Warner and Azubah Tully; m. 1850, Joel M. Gloyd (who was connected with Chief Justice Waite's family).

243311. Abigail Huntington Lathrop (widow), Boonton, N. J., b. 1810, dau. of Mary Ely and Gerard Lathrop; m. 1835, Donald McKenzie, West Indies, who was b. 1812, d. 1839.

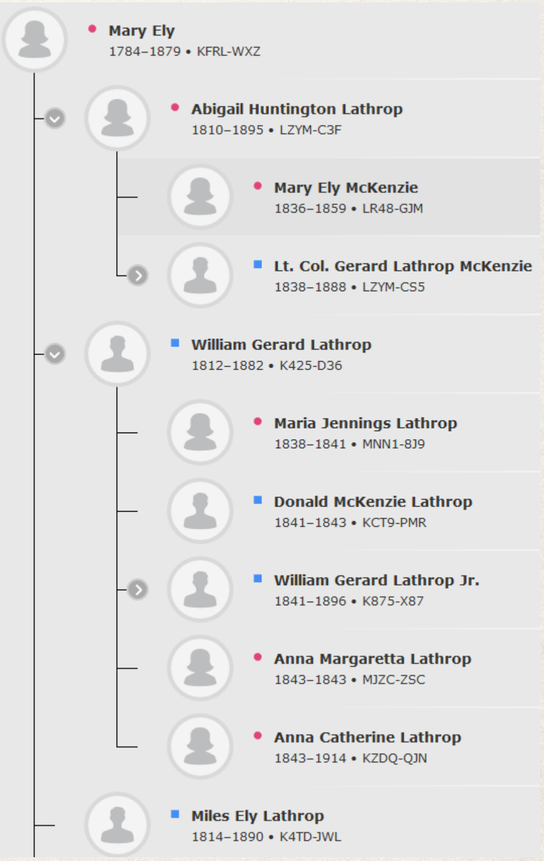
(The widow is unable to give the names of her husband's parents.) Their children:

1. Mary Ely, b. 1836, d. 1859.
2. Gerard Lathrop, b. 1838.

243312. William Gerard Lathrop, Boonton, N. J., b. 1812, d. 1882, son of Mary Ely and Gerard Lathrop; m. 1837, Charlotte Brackett Jennings, New York City, who was b. 1818, dau. of Nathan Tilestone Jennings and Maria Miller. Their children:

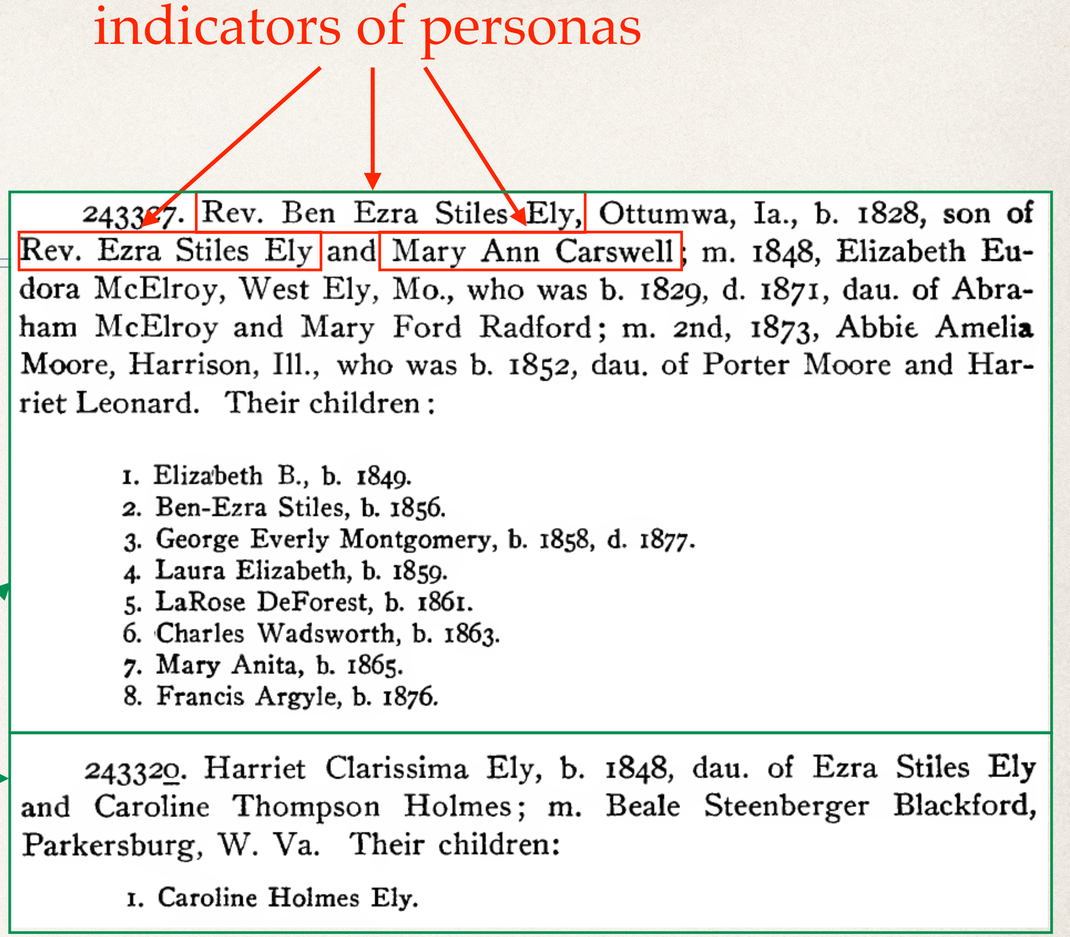
1. Maria Jennings, b. 1838, d. 1840.
2. William Gerard, b. 1840. } Twins.
3. Donald McKenzie, b. 1840, d. 1843. }
4. Anna Margaretta, b. 1843.
5. Anna Catherine, b. 1845.

243314. Charles Christopher Lathrop, N. Y. City, b. 1817, d. 1865, son of Mary Ely and Gerard Lathrop; m. 1856, Mary Augusta Andruss, 003 Broad St., Newark, N. J., who was b. 1825, dau. of Judge Caleb



Persona Records

- In a historical document, each mention of a person is a *persona*
- A *persona record* is a persona and its related information
  - For the persona "Rev. Ben Ezra Stiles Ely" we find information about birth, parents, spouses, children, etc.



Snippet from *The Ely Ancestry* (Boston, Mass., ~1650-1900)

Information Variety

- Historical documents come in many formats

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 Gilbert 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

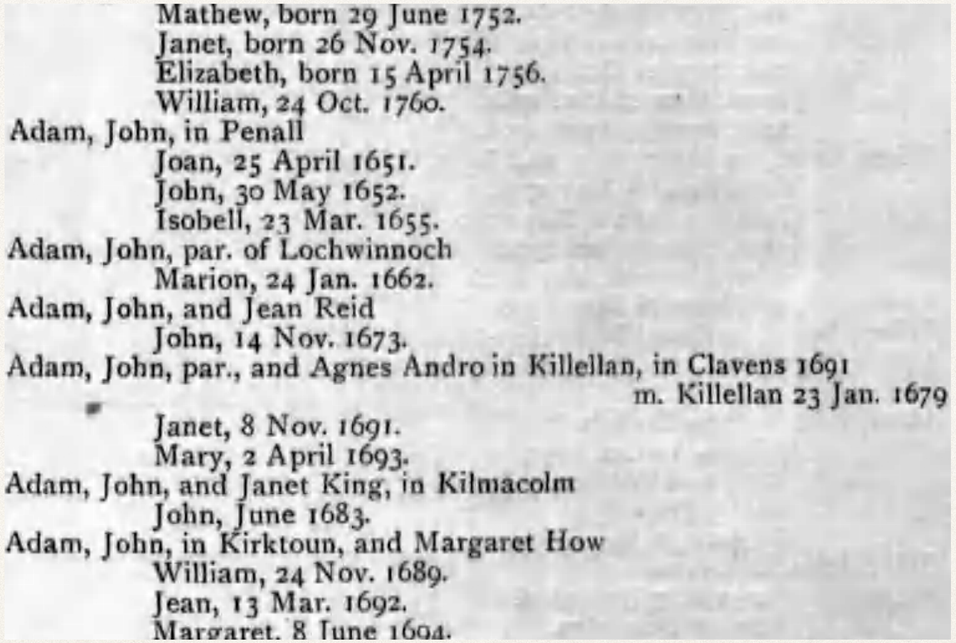
TEEGARDEN, LORENE d 2 Nov 1946 Washington D.C. residence for 2 years BD Greenville Cem Greenville OH 5 Nov 1946 single 3 brothers HAROLD of Washington D.C., ROLAND of Sidney OH, CHESTER of NY

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

Snippet from *Miller Funeral Home Records*, (Greenville, Ohio, ~1910-1950)

Information Variety

- Historical documents come in many formats



Mathew, born 29 June 1752.
Janet, born 26 Nov. 1754.
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, 2 April 1693.

Adam, John, and Janet King, in Kilmacolm
John, June 1683.

Adam, John, in Kirktoun, and Margaret How
William, 24 Nov. 1689.
Jean, 13 Mar. 1692.
Margaret, 8 June 1694.

Snippet from *Parish of Kilbarchan* (Scotland, ~1640-1780)

Information Variety

- Historical documents come in many formats

B024
Casten Böse, Brinksitzer, * ca. 1644, □ 11.4.1727, 83 J.,
∞ **An N.N.**, * ca. 1637, □ 6.5.1719, 81 ½ J.,
Kinder: Hans * ca. 1674, ∞I. 20.10.1704, B025
Rathje * ca. 1660, ∞ Ringstedt 13.6.1682 Gretje Holweges, Köhlen,
(ziehen nach Kührstedt)

B025
Hans Böse, Brinksitzer, Hs. Nr. 46, * ca. 1674 aus B024, + 10.2.1752, 78 J., in einem Graben ertrunken,
∞I. 20.10.1704 **Könke Ütjen/Itken/Itjen**, (To. Claus Ü., Brinksitzer, nachweisbar 1677), * ca. 1673, + ca. 1715, 42 J.,
Kinder: Casten ~ 25.10.1705, □ 9.11.1705
Casten ~ 28.1.1707
Claus ~ 23.9.1710,
∞II. 19.7.1716 **Anne von Soosten**, (verm. To. Johann, nicht Claus v. S., Lintig, wie im KB reg.), * 6.8.1688, + 12.3.1739,
Kinder: Anne ~ 13.6.1717
Johann ~ 15.3.1720
Hinrich ~ 12.2.1723, ∞ 20.10.1746, B027
Hans ~ 5.3.1726, ∞ 3.12.1756, B028
Stopher * 17.4.1729, + 5.2.1731
Mette Alheit * 17.8.1731, ∞ 20.11.1758 Johann Reyels, R033

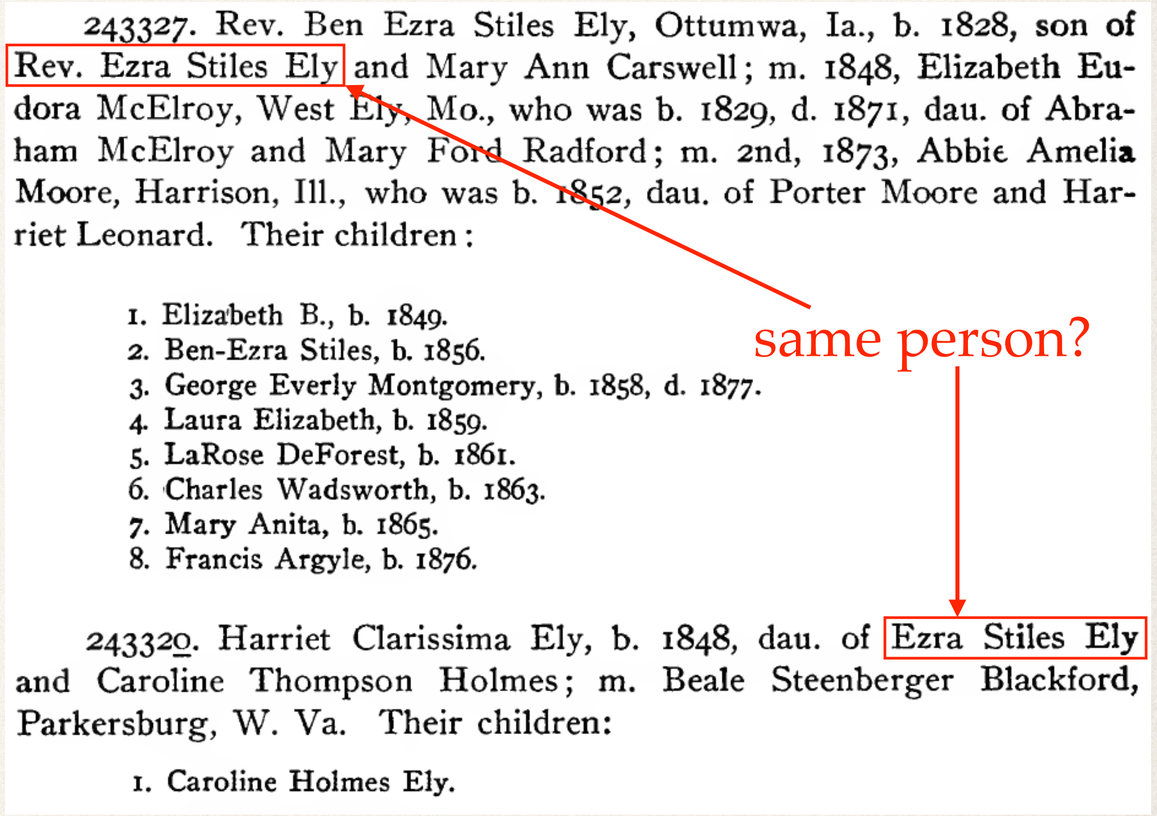
Snippet from *Familienbuch des Kirchspiels Flögeln*
(Lower Saxony, Germany, ~1670-1900)

Challenge: Linking Intergenerational Records

- Given a set of persona records, what are their intergenerational linkages?
- Requires:
  - Discovery of parent-child relationships among personas
  - Discovery of which personas refer to the same person

Example of Finding Relationships

- Two personas match in spite of having different spouses and children born 20 years apart
- Question: is there a persona elsewhere in this book that is fully consistent with the proposition that these two are the same?



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.

same person?

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.

Example of Finding Relationships

- Despite name variations, these two personas match

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 Gilbert 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

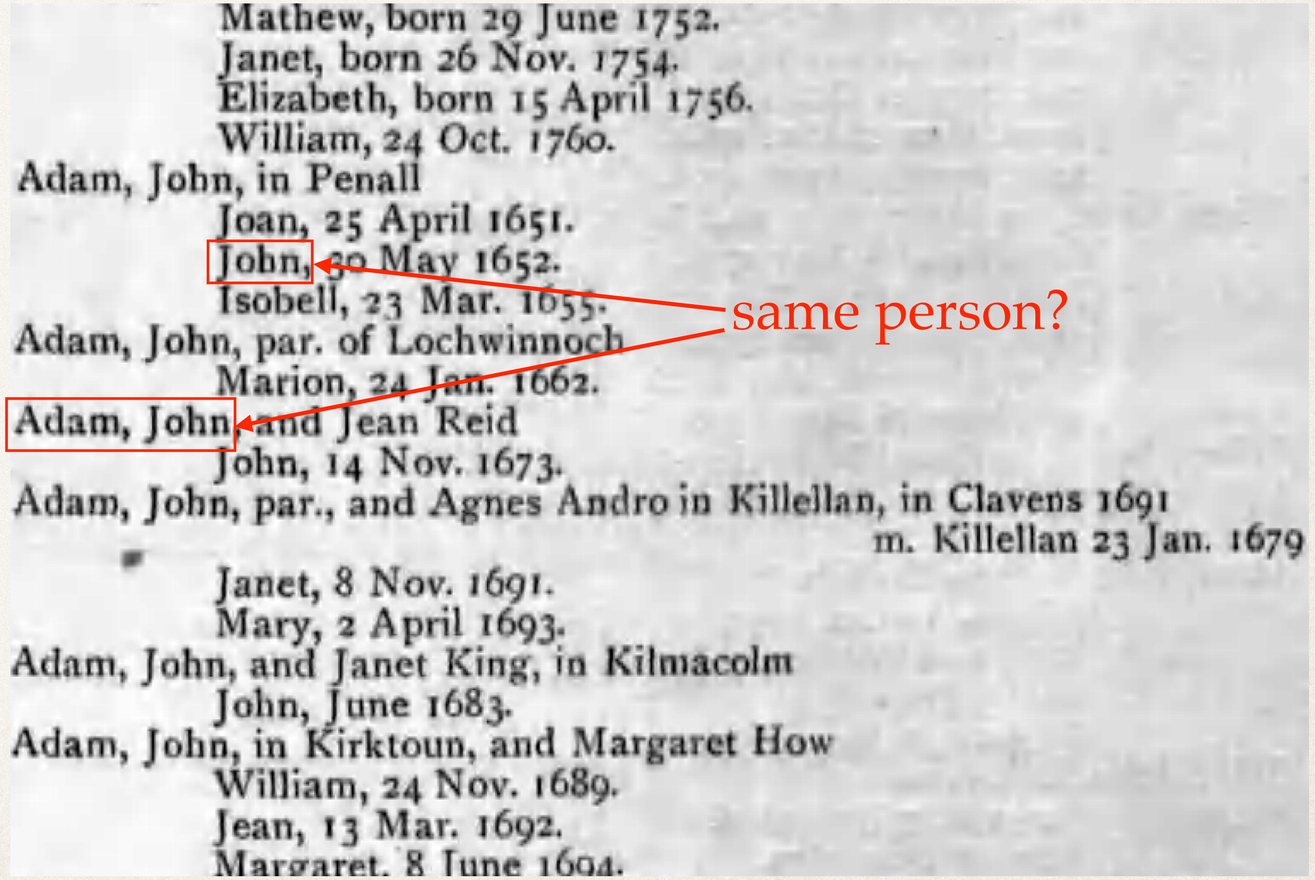
same person?

TEEGARDEN, LORENE d 2 Nov 1946 Washington D.C. residence for 2 years BD Greenville Cem Greenville OH 5 Nov 1946 single 3 brothers HAROLD of Washington 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

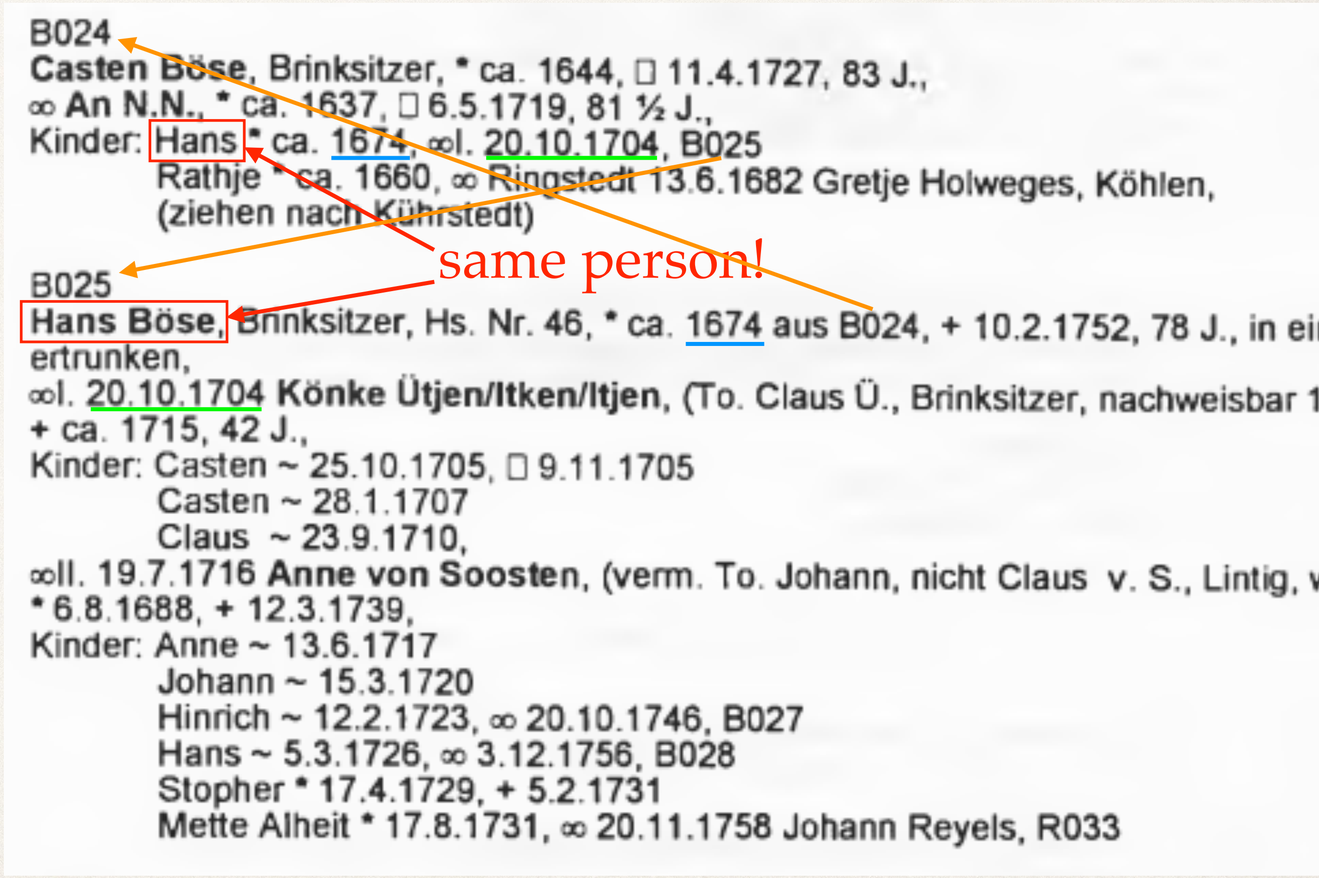
Example of Finding Relationships

- Are these two "John Adam" personas the same person?
  - $John_1$ would have been about 21 when Jean's son was born (a plausible age for a first-time father)
- There are *MANY* possible plausible connections



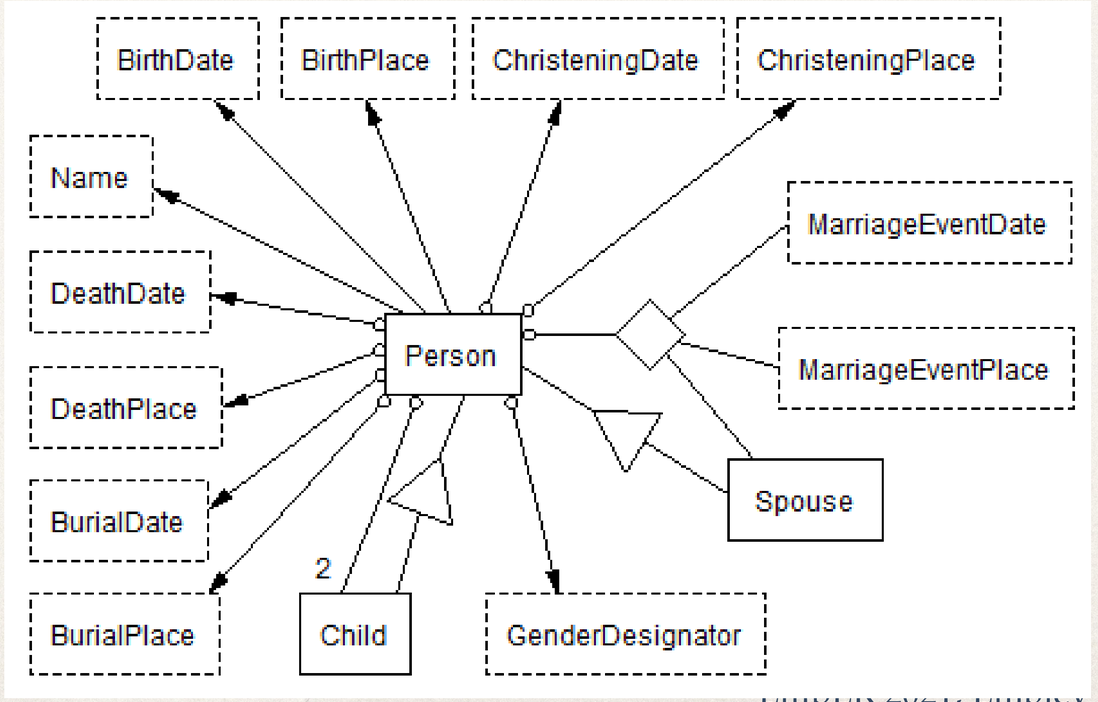
Example of Finding Relationships

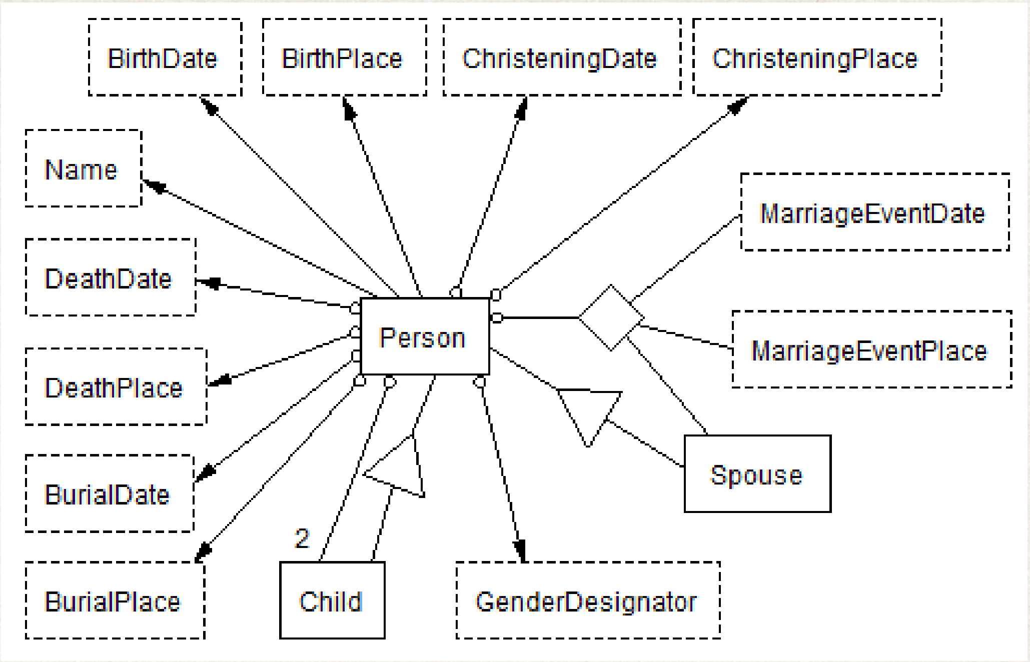
- It is clear in this case that Hans in household B024 is the same as Hans Böse in household B025
  - Not only do marriage dates match, but also household references and other facts align



OntoLink: Enriched Ontologies

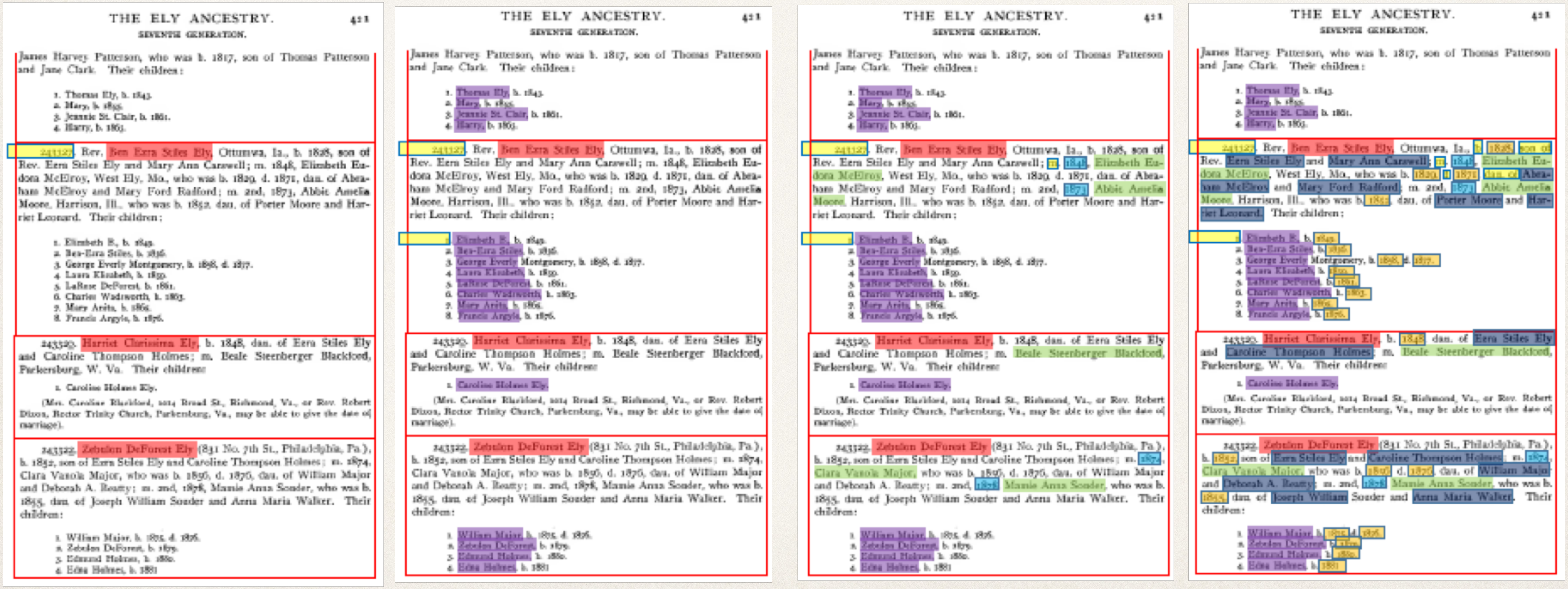
- Our system, OntoLink, uses enriched ontologies to help with the record linking task
- "An *ontology* is a formal, explicit specification of a shared conceptualization" [Gruber93]
- Conceptual model
- Enrichments:
  - Linguistic grounding
  - Pragmatic constraints
  - Cultural normatives
  - Evidential reasoning





Linguistic Grounding

(syntactically extract text elements into conceptual components)



THE ELY ANCESTRY. 421

SEVENTH GENERATION.

James Harvey Patterson, who was b. 1817, son of Thomas Patterson and Jane Clark. Their children:

1. Thomas Ely, b. 1843.
2. Mary, b. 1845.
3. Jennie St. Clair, b. 1861.
4. Harry, b. 1863.

241127. 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 E., 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. 1864.
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.

(Mrs. Caroline Blackford, 2014 Broad St., Richmond, Va., or Rev. Robert Dixon, Rector Trinity Church, Parkersburg, Va., may be able to give the date of marriage).

243322. Zebulon DeForest Ely (831 No. 7th St., Philadelphia, Pa.), b. 1852, son of Ezra Stiles Ely and Caroline Thompson Holmes; m. 1874, Clara Vanola Major, who was b. 1856, d. 1876, dau. of William Major and Deborah A. Beatty; m. 2nd, 1878, Mamie Anna Souder, who was b. 1855, dau. of Joseph William Souder and Anna Maria Walker. Their children:

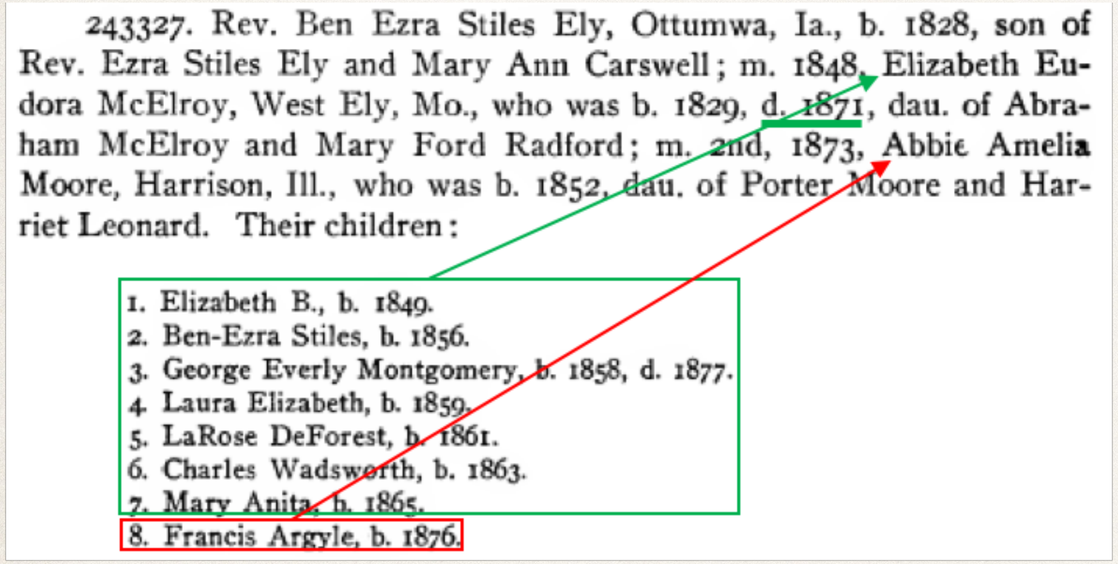
1. William Major, b. 1875, d. 1876.
2. Zebulon DeForest, b. 1879.
3. Edmund Holmes, b. 1880.
4. Edna Holmes, b. 1881.

GreenQQ Acknowledgement: George Nagy, RPI

Pragmatic Constraints

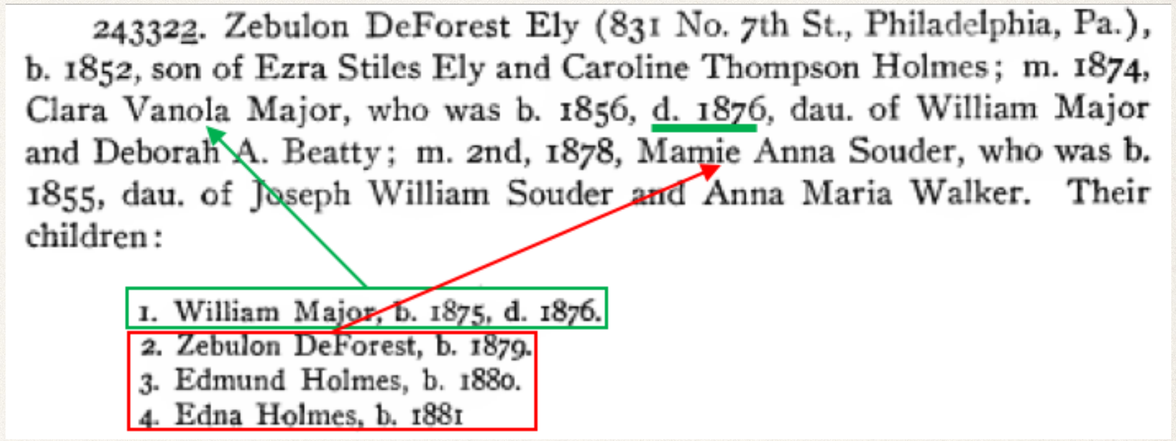
(semantic analysis of syntactically extracted information)

Example, a mother cannot give birth to a child after she dies:



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.



243322. Zebulon DeForest Ely (831 No. 7th St., Philadelphia, Pa.), b. 1852, son of Ezra Stiles Ely and Caroline Thompson Holmes; m. 1874, Clara Vanola Major, who was b. 1856, d. 1876, dau. of William Major and Deborah A. Beatty; m. 2nd, 1878, Mamie Anna Souder, who was b. 1855, dau. of Joseph William Souder and Anna Maria Walker. Their children:

1. William Major, b. 1875, d. 1876.
2. Zebulon DeForest, b. 1879.
3. Edmund Holmes, b. 1880.
4. Edna Holmes, b. 1881

Example, a person cannot die before being born: John Adams (1756 – i797)

Pragmatic Constraints

(semantic analysis of syntactically extracted information)

If a potential merge of personas violates pragmatic constraints,
OntoLink raises a red flag and rejects the merge

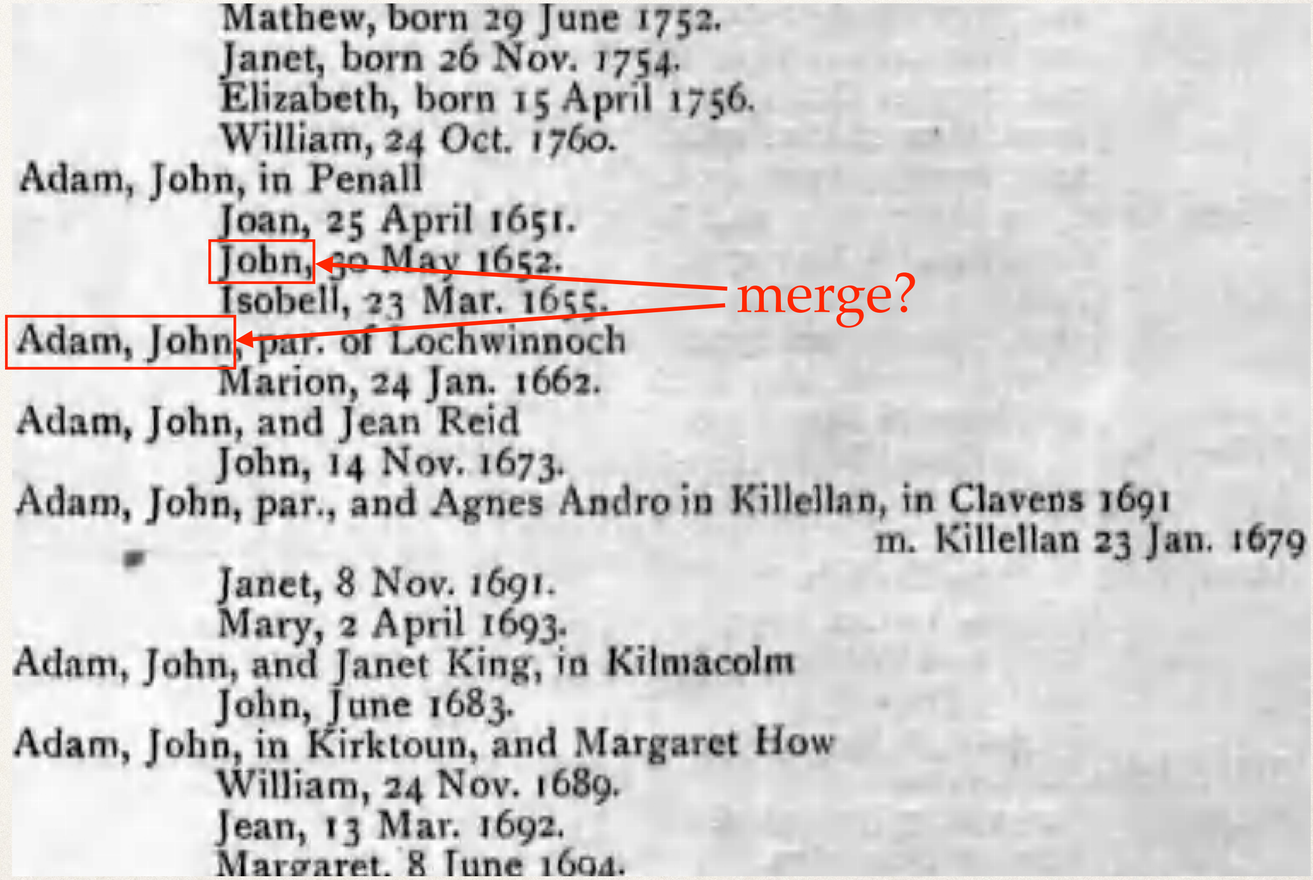
E.g. if a merge were to produce more than two parents, red flag => reject

Pragmatic Constraints

(semantic analysis of syntactically extracted information)

OntoLink also flags *improbability* (not just *impossibility*)

E.g. if we merged these two personas, it would mean that Marion, christened 24 Jan. 1662, would have been born when John was 10 years old — *improbable!*

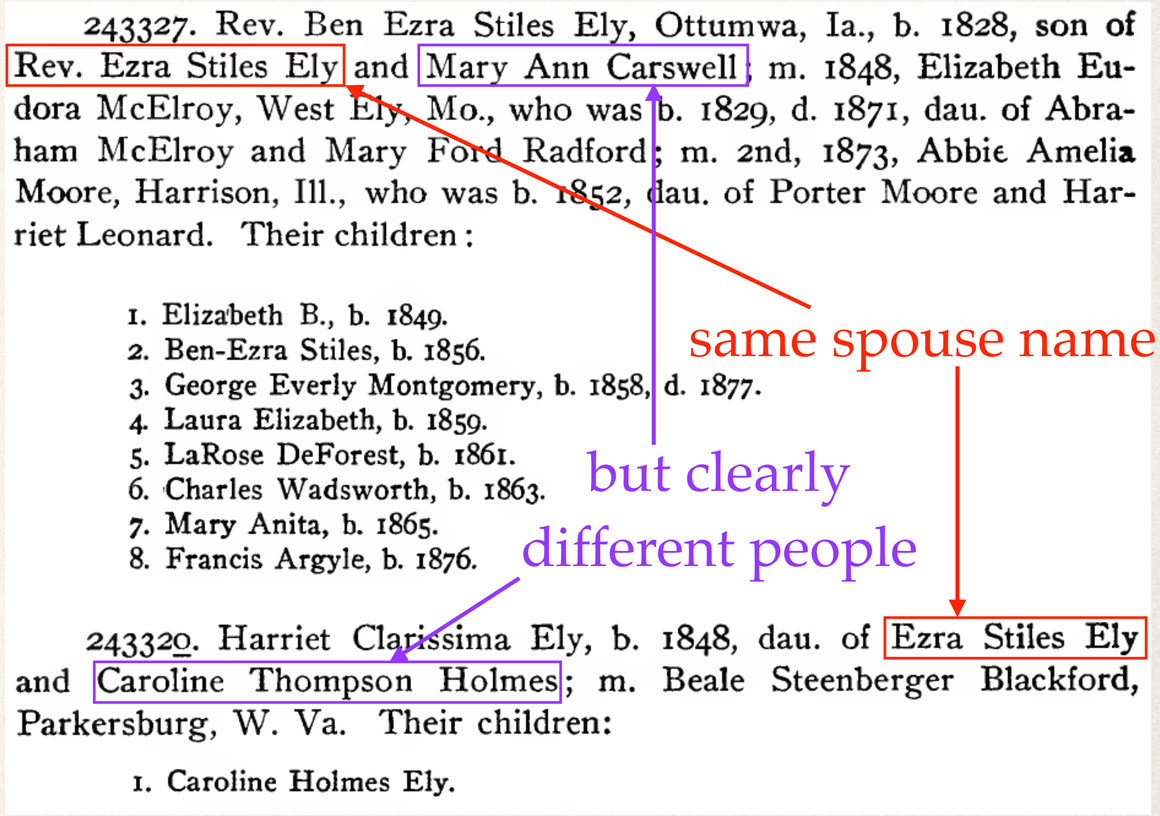


Pragmatic Constraints

(semantic analysis of syntactically extracted information)

OntoLink also flags merges where corresponding values are not close enough

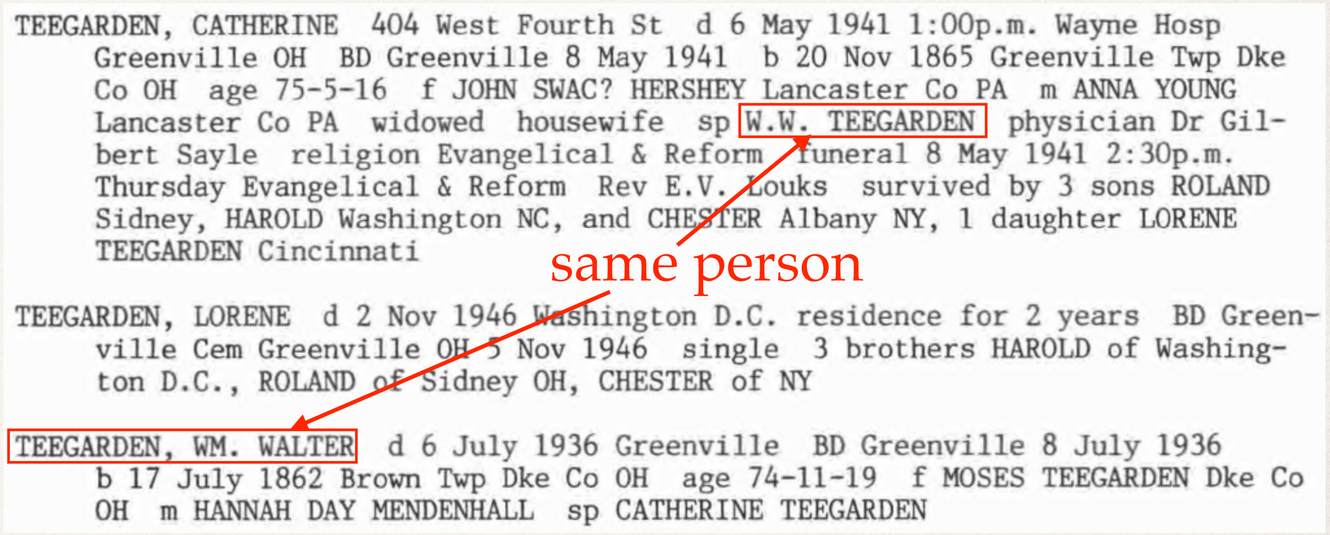
Example: *Mary Ann Carswell* and *Caroline Thompson Holmes* both married an *Ezra Stiles Ely*, but these women are clearly not the same person



Pragmatic Constraints

(semantic analysis of syntactically extracted information)

However, *W.W. TEEGARDEN* and *TEEGARDEN, WM. WALTER* do match despite different spellings and name-part orderings



Cultural Normatives

(augment extracted information by inference)

We can infer a great deal if we "read between the lines"

If we understand cultural normatives, we can infer surnames of children (same as father's surname, *Ely*, in this context)

Likewise, *TEEGARDEN* is *CATHERINE*'s married surname (her maiden surname is *HERSHEY*)

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 Gilbert 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

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.

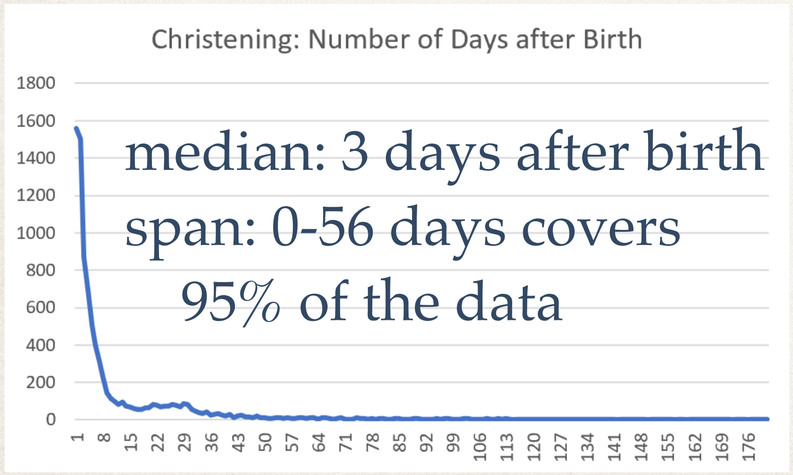
243320. Harriet Clarissima Ely, b. 1848, dau. of Ezra Stiles Ely and Caroline Thompson Holmes; m. Beale Steenberger Blackford, Parkersburg, W. Va. Their children:

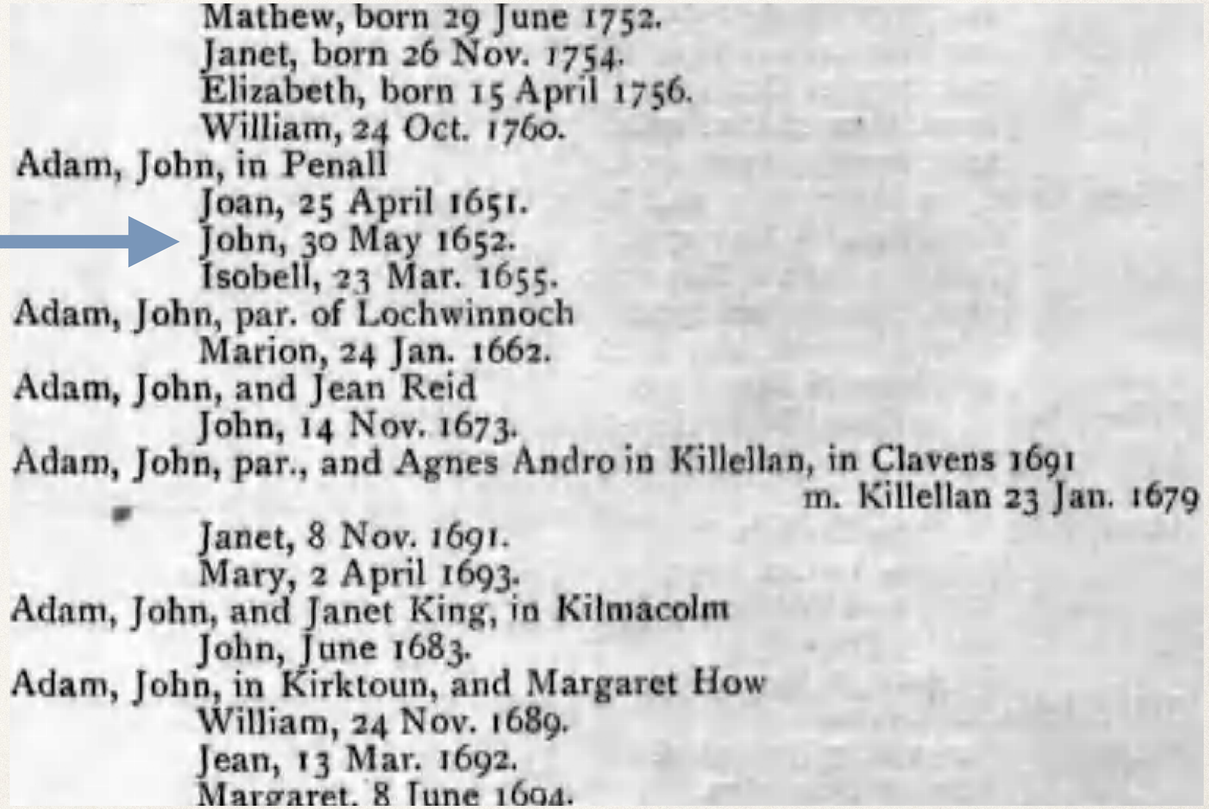
1. Caroline Holmes Ely.

Cultural Normatives

(augment extracted information by inference)

Cultural normatives and local religious practice strongly indicate that John Adam's missing date of birth was a few weeks prior to his christening date





Mathew, born 29 June 1752.
Janet, born 26 Nov. 1754.
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, 2 April 1693.

Adam, John, and Janet King, in Kilmacolm
John, June 1683.

Adam, John, in Kirktoun, and Margaret How
William, 24 Nov. 1689.
Jean, 13 Mar. 1692.
Margaret, 8 June 1694.

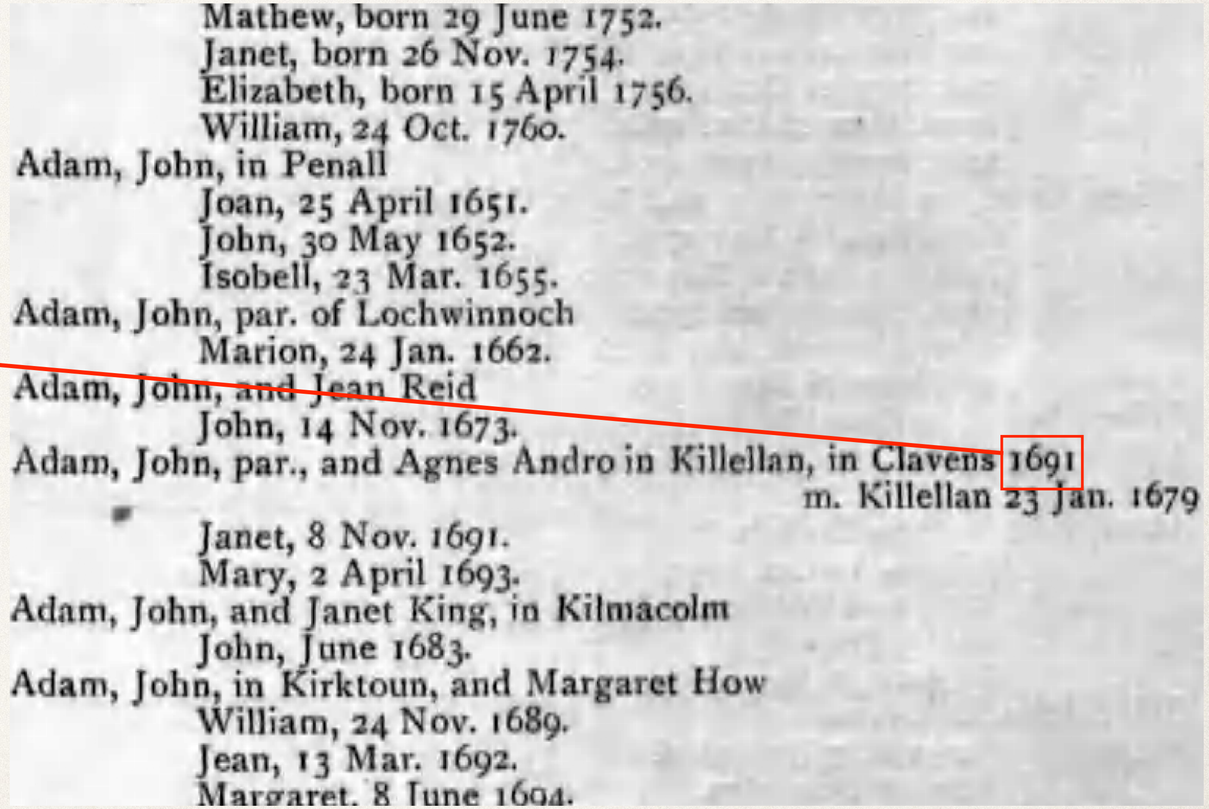
Cultural Normatives

(augment extracted information by inference)

OntoLink standardizes the text and canonicalizes the text values to ease processing

E.g. dates are canonicalized as Julian date ranges like *1691001-1691365* for the date 1691

Locations and names are also standardized and canonicalized


Mathew, born 29 June 1752.
Janet, born 26 Nov. 1754.
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, 2 April 1693.
Adam, John, and Janet King, in Kilmacolm
John, June 1683.
Adam, John, in Kirktoun, and Margaret How
William, 24 Nov. 1689.
Jean, 13 Mar. 1692.
Margaret, 8 June 1694.

Evidential Reasoning

- We need evidential reasoning to determine which persona records to merge
  - This is a record linkage problem
- Typical approach to record linkage involves:
  - Input preparation
  - Blocking
  - Within-block matching

Evidential Reasoning

- We need evidential reasoning to determine which persona records to merge
  - This is a record linkage problem
- Typical approach to record linkage involves:
  - Input preparation —more extensive
  - Blocking —governed by shallow matching on mostly inferred info
  - Within-block matching —final matching is deep based on ontological knowledge

Ontological enrichments enhance each phase

Evidential Reasoning: Input Preparation

- Persona record has a lot of information:
  - Extracted and standardized name, date, and location facts for birth, marriage, and death events
  - Extracted and inferred “one-hop” family relationships to parents, spouses, and children
- Every lexical value is standardized (to aid identity matching) and canonicalized (to aid in measured closeness matching)

Evidential Reasoning: Blocking

(a.k.a. shallow-match equivalence class construction)

- Order persona records by description information richness (most to least)
- Form ordered list of equivalence classes with each equivalence class also ordered by persona-record richness
  - The relationship for the equivalence class is: “is a plausible match with the first persona placed in the equivalence class”
  - See our paper for details of what constitutes “plausible” (names within specified edit distance, overlapping birth dates, etc.)

Evidential Reasoning: Matching

(a.k.a. deep-match equivalence class construction)

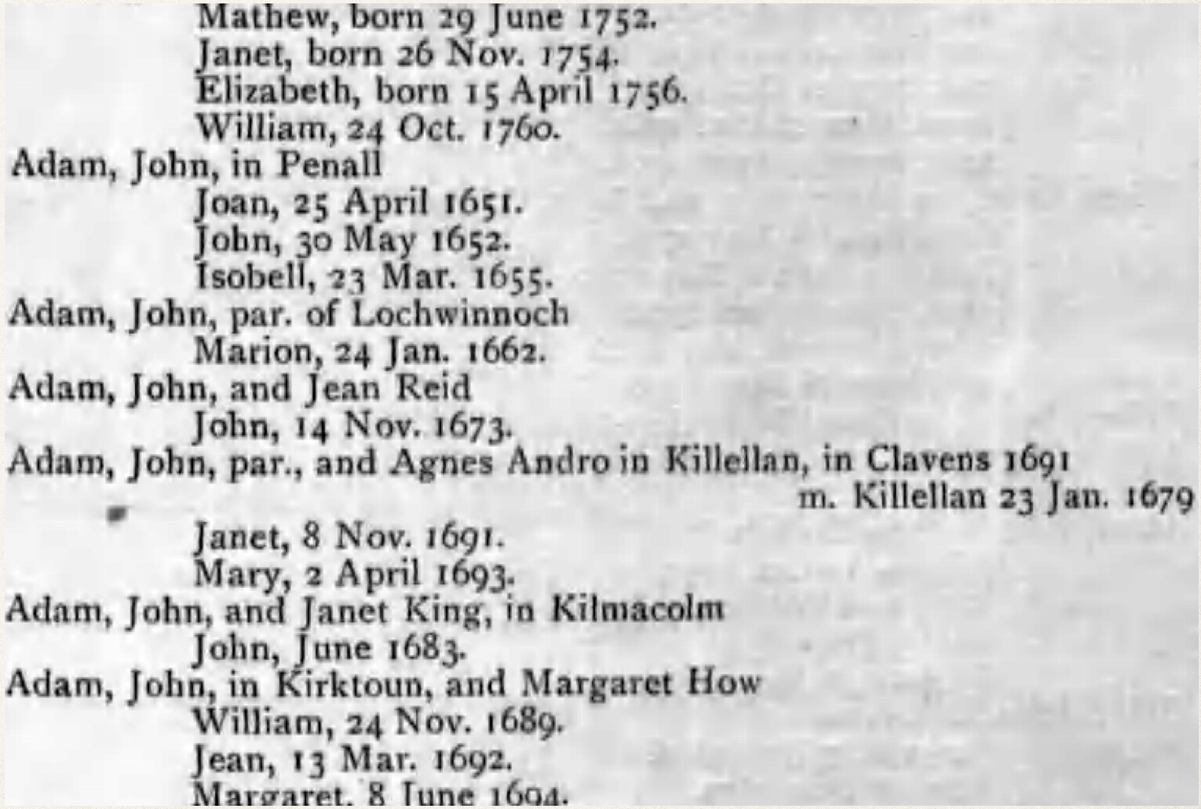
- The equivalence class relationship is: “is a match”
- The “match” check is deep and based on two ideas:
  - If two personas are merged, the merge makes sense semantically
  - The evidence for the match is sufficient to yield a high level of certainty
- When checking a shallow-match equivalence class for deep match, OntoLink compares each persona pairwise with ***all*** prior personas (to ensure that the match relationship is reflexive, symmetric, and transitive)

See the paper for details of “semantically reasonable” and “high level of certainty”

Evidential Reasoning: Matching

(a.k.a. deep-match equivalence class construction)

- Example: *Kilbarchan Parish*
- Some of the *John Adam*s shallow match, but none deep match



Mathew, born 29 June 1752.
Janet, born 26 Nov. 1754.
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, 2 April 1693.

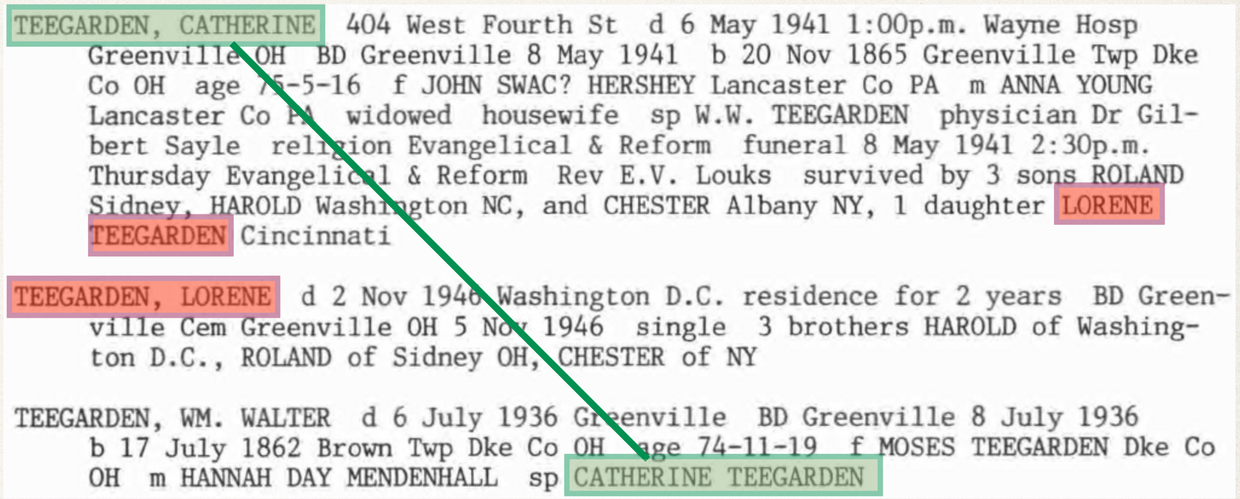
Adam, John, and Janet King, in Kilmacolm
John, June 1683.

Adam, John, in Kirktoun, and Margaret How
William, 24 Nov. 1689.
Jean, 13 Mar. 1692.
Margaret, 8 June 1694.

Evidential Reasoning: Matching

(a.k.a. deep-match equivalence class construction)

- Example: *Miller Funeral Home*
- OntoLink matches the *CATHERINE TEEGARDEN* personas (correctly)
- But (incorrectly) fails to match the *LORENE TEEGARDEN* personas
  - These personas shallow match because brothers align, but OntoLink only considers one-hop relationships, so it misses a vital clue (she has the same parents as her brothers)



Evidential Reasoning: Matching

(a.k.a. deep-match equivalence class construction)

- Example: *The Ely Ancestry*
- All three personas with *Ezra Stiles Ely* in their name shallow match, but *Rev. Ben Ezra Stiles Ely* fails the deep-match constraint that father and son cannot be the same person
  - OntoLink (correctly) places *Rev. Ben Ezra Stiles Ely* in a separate equivalence class

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.

Evidential Reasoning: Matching

(a.k.a. deep-match equivalence class construction)

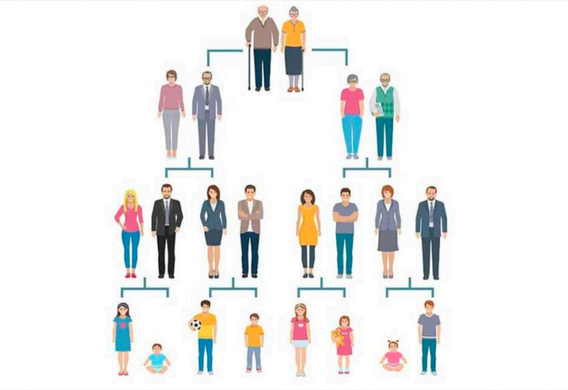
- Example: *Flögeln*
- Resolving the references boosts the match confidence of *Hans* in household *B024* and *Hans Böse* in household *B025* because marriage dates and spouse names match identically

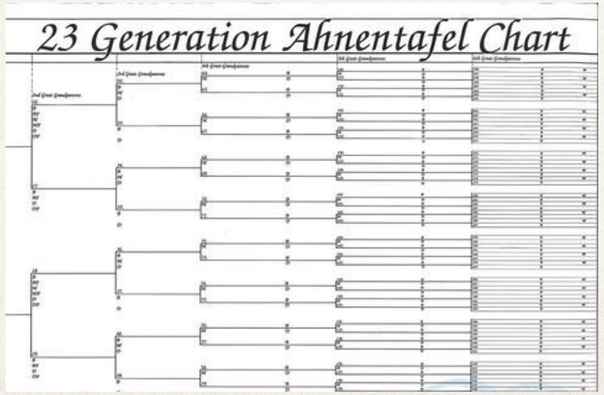
B024
Casten Böse, Brinksitzer, * ca. 1644, □ 11.4.1727, 83 J.,
∞ **An N.N.**, * ca. 1637, □ 6.5.1719, 81 ½ J.,
Kinder: Hans * ca. 1674, ∞l. 20.10.1704, B025
Rathje * ca. 1660, ∞ Ringstedt 13.6.1682 Gretje Holweges, Köhlen,
(ziehen nach Kührstedt)

B025
Hans Böse, Brinksitzer, Hs. Nr. 46, * ca. 1674 aus B024, + 10.2.1752, 78 J., in ei
ertrunken,
∞l. 20.10.1704 **Könke Ütjen/Itken/Itjen**, (To. Claus Ü., Brinksitzer, nachweisbar 1
+ ca. 1715, 42 J.,
Kinder: Casten ~ 25.10.1705, □ 9.11.1705
Casten ~ 28.1.1707
Claus ~ 23.9.1710,
∞II. 19.7.1716 **Anne von Soosten**, (verm. To. Johann, nicht Claus v. S., Lintig, v
* 6.8.1688, + 12.3.1739,
Kinder: Anne ~ 13.6.1717
Johann ~ 15.3.1720
Hinrich ~ 12.2.1723, ∞ 20.10.1746, B027
Hans ~ 5.3.1726, ∞ 3.12.1756, B028
Stopher * 17.4.1729, + 5.2.1731
Mette Alheit * 17.8.1731, ∞ 20.11.1758 Johann Reyels, R033

Inter-Generational Family Tree Generation

- Personas in a deep-match equivalence class are guaranteed to be mergeable
- Merged personas have all the information needed to display an inter-generational family tree as pedigree chart, German *Ahnentafel*, Chinese *Jiapu*, or any other desired form





Experimental Results

- We ran OntoLink on *Ely*, *Miller*, *Kilbarchan*, *Flögeln*
  - There were 34,553 persona records
  - OntoLink inferred birth surnames for 31%, married surnames for 26%
  - For birth dates, 34% extracted, 58% estimated, 7% unknown
  - Shallow-match blocking generated 26,063 shallow equivalence classes

Experimental Results

- Persona record shallow match equivalence classes

Book (pages)	# persona records	surnames inf.		birth dates		# equiv. cls. by size		
		birth	married	extracted	estimated	1	2–9	10^+
Ely (432–700)	8,976	2,731	3,038	4,427	3,895	5,415	1,208	8
Miller (7–395)	11,439	1,532	2,573	2,818	8,303	7,749	1,554	1
Kilbarchan (4–127)	8,814	4,043	2,064	1,103	6,224	5,049	1,174	15
Flögeln (11–121)	5,324	2,557	1,382	3,451	1,771	2,924	965	1

Experimental Results: Matching Phase

- There were 31,119 deep-match equivalence classes
- OntoLink red-flagged 588 personas as not self-consistent and pushed 37,287 downstream for further processing (sometimes multiple times)
- 14,217 were unmergeable, 857 were red-flagged when merged, 22,213 were unconfident (lacking sufficient evidence to merge)
- Because red-flag violations are based on ontologically specified constraints, each red-flag violation constitutes an explanation about why two persona records cannot be merged
- For unconfident merges, we can generate a research plan for resolving merge questions
  - Yellow/green-flag probabilities associated with pragmatic constraints, cultural normatives, and relative attribute weights indicating importance of each kind of missing information can guide

Experimental Results

- Persona record deep match equivalence classes

Book (pages)	# equiv. cls. by size 1	2–9	10^+	# redflgd	# pushed downstream unmrgbl	merge redflgd	not confident
Ely (432–700)	6,479	865	2	146	3,312	1	3,615
Miller (7-395)	10,164	572	41	0	2,092	5	6,493
Kilbarchan (4–127)	8,334	12	0	438	7,819	0	10,955
Flögeln (11–121)	4,084	566	0	4	994	851	1,150

Benefit of Blocking before Matching

- Recall that our shallow matching is a "blocking" technique
- Shallow-match blocking is O(n)
- Deep matching is O(n^2)
- It would have taken ~5 days to do the deep matching on *Ely* without first performing shallow-match blocking
- Having done the shallow-match blocking, it actually took 145 seconds to process *Ely*

Experimental Results

- Persona deep match equivalence class accuracy

	false positives	false negatives	# checked (Accuracy)	Accuracy*		
				Recall	Precision	F-score
Ely (432–700)	2	16	80	83%	98%	90%
Miller (7-395)	9	4	80	95%	89%	92%
Kilbarchan (4–127)	12	0	8346	100%	99.86%	99.93%
Flögeln (11–121)	1	6	80 | 732	94% | 99.3%	99% | 99.9%	96% | 99.6%

*$\textbf{Recall} = tp/(tp+fn)$, $\textbf{Precision} = tp/(tp+fp)$, where tp is **t**rue **p**ositives, fp is **f**alse **p**ositives, and fn is **f**alse **n**egatives, and **F-score** is the harmonic mean of precision and recall, computed as $2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$.

Experimental Results

- Correctness depends on source documents being error free and OntoLink properly capturing and curating document-provided information
- For *Ely*, *Miller*, *Kilbarchan* we checked a random sample of three pages:
  - F-scores after Linguistic Grounding with GreenQQ were 90%, 81%, 86%
  - For Pragmatic Constraint identification/rectification were 97%, 87%, 92%
- Selected 10 random persona records for each book and found:
  - After processing Cultural Normatives, F-scores were 98%, 98%, 98%
- This highlights the value of ontological semantic enrichment

Value of a Ground Truth

- The accuracy results we reported for *Flögeln* depend on an extraction ground truth that was meticulously prepared manually
  - This approach doesn't scale
- So we did a robustness check by randomly selecting a single page (15) to prime GreenQQ and then ran the OntoLink pipeline under four conditions:
  - Using the Adobe Acrobat OCR engine
  - Using the ABBYY Fine Reader OCR engine
  - + introducing page cleaning and OCR corrections
  - + allowing GreenQQ to give active feedback to the human operator for tuning details

Experimental Results

- We measured PRF scores and then repeated with another random seed page (46)
- GreenQQ extraction accuracy for *Flögeln* under various conditions:

	Page 15			Page 46		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
No page cleaning; no OCR corrections; Adobe Acrobat OCR engine	62%	50%	56%	62%	51%	56%
No page cleaning; no OCR corrections; ABBYY FineReader 15 OCR engine	88%	79%	83%	88%	79%	83%
With page cleaning and OCR corrections; no active learning	97%	89%	93%	97%	89%	93%
With page cleaning and OCR corrections; with active learning	98%	93%	95%	97%	93%	95%

Experimental Results

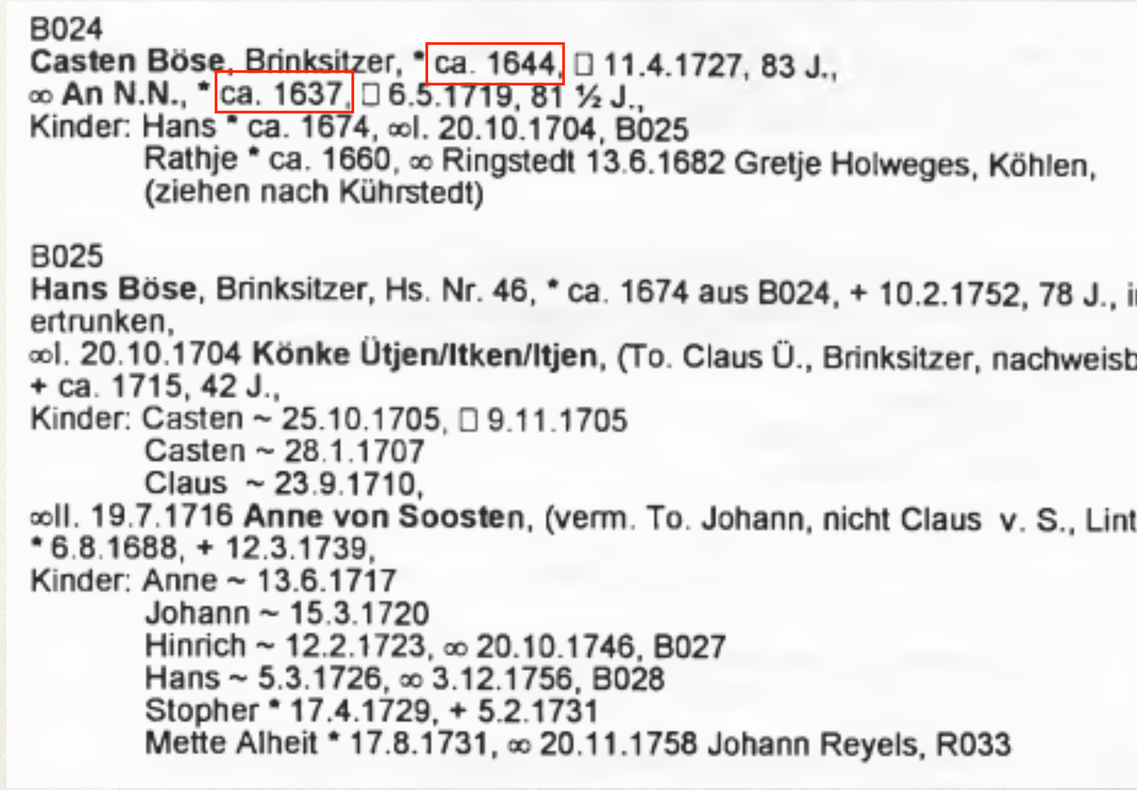
- By *active learning* we mean that GreenQQ can detect when it has failed to properly extract a record

  GreenQQ extracted two birthdates for Casten Böse, which ontologically is not allowed

  Indeed, a missing name is present, “An N.N.” which does not match a name pattern known to GreenQQ

  N.N. stands for Latin *nomen nescio*, “I do not know the name”

  When we add a template for N.N. GreenQQ can extract the proper name, “An”



B024
Casten Böse, Brinksitzer, * ca. 1644, □ 11.4.1727, 83 J.,
∞ **An N.N.**, * ca. 1637, □ 6.5.1719, 81 ½ J.,
Kinder: Hans * ca. 1674, ∞I. 20.10.1704, B025
Rathje * ca. 1660, ∞ Ringstedt 13.6.1682 Gretje Holweges, Köhlen, (ziehen nach Kührstedt)

B025
Hans Böse, Brinksitzer, Hs. Nr. 46, * ca. 1674 aus B024, + 10.2.1752, 78 J., i
ertrunken,
∞I. 20.10.1704 **Könke Ütjen/Itken/Itjen**, (To. Claus Ü., Brinksitzer, nachweisb
+ ca. 1715, 42 J.,
Kinder: Casten ~ 25.10.1705, □ 9.11.1705
Casten ~ 28.1.1707
Claus ~ 23.9.1710,
∞II. 19.7.1716 **Anne von Soosten**, (verm. To. Johann, nicht Claus v. S., Lint
* 6.8.1688, + 12.3.1739,
Kinder: Anne ~ 13.6.1717
Johann ~ 15.3.1720
Hinrich ~ 12.2.1723, ∞ 20.10.1746, B027
Hans ~ 5.3.1726, ∞ 3.12.1756, B028
Stopher * 17.4.1729, + 5.2.1731
Mette Alheit * 17.8.1731, ∞ 20.11.1758 Johann Reyels, R033

Experimental Results

- Creating a ground truth can be quite hard, but the OntoLink pipeline performs quite well even with a much less intensive process of:
  - Page cleaning
  - OCR correction
  - Markup of one page of example templates
  - Iterative feedback to prompt for additional example templates to resolve obviously incorrect extractions

Experimental Results

- Deep-match equivalence classes comprise a book's family trees
- Deep-match equivalence class F-scores ranged from 90% to 99%
- OntoLink was able to automatically create inter-generational family trees with accuracy in the 90th percentile

Future Work

- 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
  - Additional testing to adjust constraints, fine-tune parameters and thresholds

Conclusion

- The results of this preliminary study are promising
- Evidence supports the claim that enriching an ontology with linguistic grounding, pragmatic constraints, cultural normatives, and evidential reasoning can measurably enhance the work of record linkage
  - A contribution within the digital humanities