Near-perfect Relation Extraction

semi-structured

from ^ Family Books

George Nagy
Rensselaer Polytechnic Institute
Troy, NY USA

Slides for remote participants

Near perfect!

ne plus ultra?

3 books, ~ 500,000 word tokens ⟹ 55,900 relations

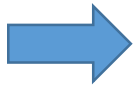
> 99.5 % precision, recall, F-measure

Suitable only for semi-structured text, good OCR, selected binary relations, English text, observant user. For other restrictions see verso.

Marie, daughter of James, married Martin, son of Daniel, in Boston

Named Entities

Marie	Head
James	Father
Martin	Spouse
Daniel	Father



Relations

Head[Spouse]	Marie[Martin]
Father[Head]	James[Marie]
Father[Spouse]	Daniel[Martin]
Father[Head[Spouse]]	Daniel[Marie[Martin]]

“Family” Books

>300,000 digitized books, pamphlets and records of genealogical data await processing.

Family Books are lists of quasi-repetitive records where some tokens can be designated as *signs* or *values* of the information to be extracted.

OCR’d family book → Unicode page text

243353. Edwin Clark Hyde (13 Warren St., St. Louis), b. 1819,
son of Julia Ely and Zabdial Hyde ; m. 1844, Elizabeth Ann Peake (Gor-
don), who was b. 1816, dau. of Henry Peake and Isabella Herring
(Snyder) . Their children :
1. Henrietta Mills, b. 1845, d. 1848.
2. Susan Isabella, b. 1847.
3. Samuel Peake, b. 1850.
4. Annie Carroll, b. 1851, d. 1857.
5. Allen Withers, b. 185S, d. 1856.

OCR error 1855? 1858?

Semi-structure

Signs (*born* or *son of*) and *values* (*1828* or *Henrietta Mills*) alternate.

Semi-structure can often be imposed on family books
by good choice of signs and values.

Page text → family records (HIP 2021)

…HEAD:,578,10,9,3,Henrietta,Mills,Hyde,B_DATE:,578,11,9,1,1826,PARENT1:,
578,11,23,2,Julia,Ely,PARENT2:,578,11,37,2,Zabdal,Hyde,M_DATE:,578,11,54,1,
1848,SPOUSE:,578,11,60,3,Charles,Smith,Shelton,B_DATE:,578,12,47,1,1819,
D_DATE:,578,12,56,1,1879,PARENT1:,578,13,1,2,George,Shelton,PARENT2:,
578,13,20,2,Betsy,Wooster,CHILD:,578,14,4,2,Fany,Arabella,B_DATE:,578,14,23,1,
1850,M_DATE:,578,14,32,1,1874,SPOUSE:,578,14,38,3,Arthur,Harry,Bissell,
CHILD:,578,15,4,2,Julia,Elizabeth,B_DATE:,578,15,24,1,1851,M_DATE:,578,15,34
,1,1878,SPOUSE:,578,15,40,5,Chas,.,J.,Van,Tsel,CHILD:,578,16,4,2,Charles,Henry,
B_DATE:,578,16,22,1,1854,CHILD:,578,17,4,2Henry,Hyde,B_DATE:,578,17,19,1,
1858,HEAD:, …

class page line offset length extract

…..,PARENT2:,578,11,37,2,Zabdal,Hyde……

Page text → family records (HIP 2021)

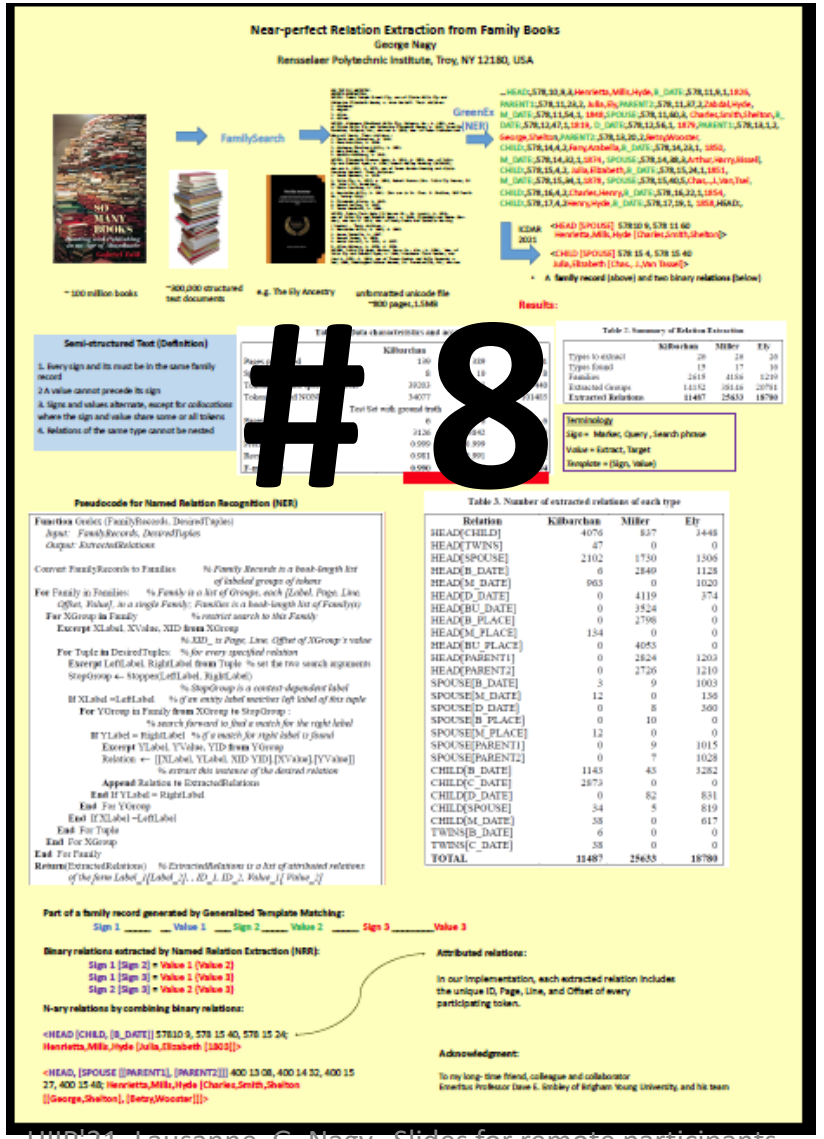
…HEAD:,578,10,9,3,Henrietta,Mills,Hyde,B_DATE:,578,11,9,1,1826,PARENT1:,
578,11,23,2,Julia,Ely,PARENT2:,578,11,37,2,Zabdal,Hyde,M_DATE:,578,11,54,1,
1848,SPOUSE:,578,11,60,3,Charles,Smith,Shelton,B_DATE:,578,12,47,1,1819,
D_DATE:,578,12,56,1,1879,PARENT1:,578,13,1,2,George,Shelton,PARENT2:,
578,13,20,2,Betsy,Wooster,CHILD:,578,14,4,2,Fany,Arabella,B_DATE:,578,14,23,1,
1850,M_DATE:,578,14,32,1,1874,SPOUSE:,578,14,38,3,Arthur,Harry,Bissell,
CHILD:,578,15,4,2,Julia,Elizabeth,B_DATE:,578,15,24,1,1851,M_DATE:,578,15,34
,1,1878,SPOUSE:,578,15,40,5,Chas,.,J.,Van,Tsel,CHILD:,578,16,4,2,Charles,Henry,
B_DATE:,578,16,22,1,1854,CHILD:,578,17,4,2Henry,Hyde,B_DATE:,578,17,19,1,
1858,HEAD:, …

class page line offset length extract

…..,PARENT2:,578,11,37,2,Zabdal,Hyde……

Family records ⟶ Relations

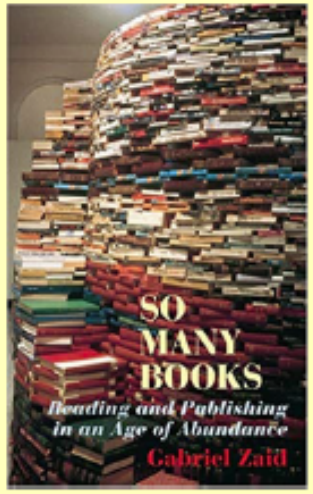
The next four slides display this poster slice by slice.



Slice # 1

Near-perfect Relation Extraction from Family Books

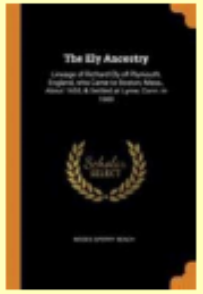
George Nagy
Rensselaer Polytechnic Institute, Troy, NY 12180, USA



~ 100 million books

FamilySearch

~300,000 structured text documents



e.g. The Ely Ancestry

422 THE ELY ANCESTRY.
SEVENTH GENERATION.
2433[illegible]- James Joseph Ernest Ely, son of Elisha Mills Ely and Catherine Elizabeth Woode; m. Anna Horloff. Their children:
1. Alphonse.
2. August.
3. Alice.
4. Alfred.
2433[illegible]. Alphonso Sthelbert Mills Ely, Palmyra, Mo., b. 1821, son of Elisha Mills Ely and Catherine Elizabeth Woode; m. 1841, Drusilla Pinkston, Palmyra, Mo., who was b. 1820, dau. of Peter Pinkston and Abig-ail Davis. Their children :
1. Laura Ann Catherine, b. 1842.
2. Emma McClellan, b. 1850.
3. Alphonse Sthelbert Mills, b. 1852.
4. Mary Bailey, b. 1856.
5. Ophelia Goldburg, b. 1861.
2433[illegible]. Elizabeth Plummer Hyde, b. 1814, d. 1855, dau. of Julia Ely and Zabdial Hyde; m. 1834, Robert McClay Henning (2433623), who was b. 1812, d. 1875, son of James Gordon Henning and Alicia Courtney Spinner. Their children:
1. James Spencer, b. 1835.
2. Julia Ely, b. 1837; m. 1856, Robert Pearce (Mrs. Julia Ely Pearce, 58 St. John's Pl., Brooklyn).
3. Edwin Courtney, b. 1838.
4. Henrietta Mills, b. 1841. (Her son is Dr. Chas. H. Shelton, 288 Fourth St., Jersey City).
5. Elizabeth Alicia, b. 1843.
6. Robert McClay, b. 1847.
7. James Woodruff, b. 1850.
2433[illegible]. Edwin Clark Hyde (13 Warren St., St. Louis), b. 1819, son of Julia Ely and Zabdial Hyde ; m. 1844, Elizabeth Ann Peake (Morton), who was b. 1826, dau. of Henry Peake and Isabella Herring (Snyder) . Their children :
1. Henrietta Mills, b. 1845, d. 1848.
2. Susan Isabella, b. 1847.
3. Samuel Peake, b. 1850.
4. Annie Carroll, b. 1851, d. 1857.
5. Allen Withers, b. 1855, d. 1856.
2433[illegible]. Julia Ely Hyde, Marion, Perry Co., Ala., b. 1824, dau. of Julia Ely and Zabdial Hyde; m. 1846, Alexander Clark Bunker, who (was b. 1822, d. 1848, son of Thomas Bunker and Sally Raymond; m. 2nd, 1854, Washington Holmes Havent, St. Francisville, Mo., who was

unformatted unicode file
~800 pages,1.5MB

GreenEx (NER)

...HEAD:,578,10,9,3,Henrietta,Mills,Hyde,B_DATE:,578,11,9,1,1826,
PARENT1:,578,11,23,2, Julia,Ely,PARENT2:,578,11,37,2,Zabdal,Hyde,
M_DATE:,578,11,54,1, 1848,SPOUSE:,578,11,60,3, Charles,Smith,Shelton,B_
DATE:,578,12,47,1,1819, D_DATE:,578,12,56,1, 1879,PARENT1:,578,13,1,2,
George,Shelton,PARENT2:,578,13,20,2,Betsy,Wooster,
CHILD:,578,14,4,2,Fany,Arabella,B_DATE:,578,14,23,1, 1850,
M_DATE:,578,14,32,1,1874, SPOUSE:,578,14,38,3,Arthur,Harry,Bissell,
CHILD:,578,15,4,2, Julia,Elizabeth,B_DATE:,578,15,24,1,1851,
M_DATE:,578,15,34,1,1878, SPOUSE:,578,15,40,5,Chas,.,J.,Van,Tsel,
CHILD:,578,16,4,2,Charles,Henry,B_DATE:,578,16,22,1,1854,
CHILD:,578,17,4,2Henry,Hyde,B_DATE:,578,17,19,1, 1858,HEAD:,

ICDAR 2021

<HEAD [SPOUSE] 57810 9, 578 11 60
Henrietta,Mills,Hyde [Charles,Smith,Shelton]>

<CHILD [SPOUSE] 578 15 4, 578 15 40
Julia,Elizabeth [Chas., J.,Van Tassel]>

- A family record (above) and two binary relations (below)

Results:

Results:

Semi-structured Text (Definition)

1. Every sign and its must be in the same family record
2 A value cannot precede its sign
3. Signs and values alternate, except for *collocations* where the sign and value share some or all tokens
4. Relations of the same type cannot be nested

Table 1. Data characteristics and accuracy

	Kilbarchan	Miller	Ely
Pages processed	139	389	301
Specified labels	8	10	8
Tokens assigned specified labels	39203	91633	39440
Tokens labeled NONE	34077	131462	101485
Test Set with ground truth			
Pages	6	6	6
Tokens (including "NONE")	3126	3842	3423
Precision	0.999	0.999	0.997
Recall	0.981	0.991	0.992
F-measure	**0.990**	**0.996**	**0.994**

Table 2. Summary of Relation Extraction

	Kilbachan	Miller	Ely
Types to extract	26	26	26
Types found	15	17	16
Families	2615	4186	1219
Extracted Groups	14152	38146	20781
Extracted Relations	**11487**	**25633**	**18780**

Terminology

Sign = Marker, Query , Search phrase

Value = Extract, Target

Template = (Sign, Value)

Slice # 3

Pseudocode for Named Relation Extraction (NRE)

```
Function Grelex (FamilyRecords, DesiredTuples)
   Input:  FamilyRecords, DesiredTuples
   Output: ExtractedRelations

Convert FamilyRecords to Families        % Family Records is a book-length list
                                           of labeled groups of tokens
For Family in Families:      % Family is a list of Groups, each [Label, Page, Line,
     Offset, Value], in a single Family; Families is a book-length list of Family(s)
  For XGroup in Family                   % restrict search to this Family
    Excerpt XLabel, XValue, XID from XGroup
                                         % XID_ is Page, Line, Offset of XGroup's value
    For Tuple in DesiredTuples:          % for every specified relation
      Excerpt LeftLabel, RightLabel from Tuple  % set the two search arguments
      StopGroup ← Stopper(LeftLabel, RightLabel)
                                         % StopGroup is a context-dependent label
      If XLabel =LeftLabel     % if an entity label matches left label of this tuple
        For YGroup in Family from XGroup to StopGroup :
                                 % search forward to find a match for the right label
          If YLabel = RightLabel  % if a match for right label is found
            Excerpt YLabel, YValue, YID from YGroup
            Relation ← [[XLabel, YLabel, XID YID],[XValue],[YValue]]
                                 % extract this instance of the desired relation
            Append Relation to ExtractedRelations
          End If YLabel = RightLabel
        End  For YGroup
      End  If XLabel =LeftLabel
    End  For Tuple
  End  For XGroup
End  For Family
Return(ExtractedRelations)   % ExtractedRelations is a list of attributed relations
     of the form Label_1[Label_2], , ID_1, ID_2, Value_1[,Value_2]
```

Table 3. Number of extracted relations of each type

Relation	Kilbarchan	Miller	Ely
HEAD[CHILD]	4076	837	3448
HEAD[TWINS]	47	0	0
HEAD[SPOUSE]	2102	1730	1306
HEAD[B_DATE]	6	2849	1128
HEAD[M_DATE]	963	0	1020
HEAD[D_DATE]	0	4119	374
HEAD[BU_DATE]	0	3524	0
HEAD[B_PLACE]	0	2798	0
HEAD[M_PLACE]	134	0	0
HEAD[BU_PLACE]	0	4053	0
HEAD[PARENT1]	0	2824	1203
HEAD[PARENT2]	0	2726	1210
SPOUSE[B_DATE]	3	9	1003
SPOUSE[M_DATE]	12	0	136
SPOUSE[D_DATE]	0	8	360
SPOUSE[B_PLACE]	0	10	0
SPOUSE[M_PLACE]	12	0	0
SPOUSE[PARENT1]	0	9	1015
SPOUSE[PARENT2]	0	7	1028
CHILD[B_DATE]	1143	43	3282
CHILD[C_DATE]	2873	0	0
CHILD[D_DATE]	0	82	831
CHILD[SPOUSE]	34	5	819
CHILD[M_DATE]	38	0	617
TWINS[B_DATE]	6	0	0
TWINS[C_DATE]	38	0	0
TOTAL	**11487**	**25633**	**18780**

Slice # 4

Part of a family record generated by Generalized Template Matching:

Sign 1 _____ __ Value 1 ___ Sign 2 _____ Value 2 _____ Sign 3 _____Value 3

Binary relations extracted by Named Relation Extraction (NRR):

Sign 1 [Sign 2] = Value 1 {Value 2}
Sign 1 [Sign 3] = Value 1 {Value 3}
Sign 2 [Sign 3] = Value 2 {Value 3}

N-ary relations by combining binary relations:

<HEAD [CHILD, [B_DATE]] 57810 9, 578 15 40, 578 15 24; Henrietta,Mills,Hyde [Julia,Elizabeth [1803]]>

<HEAD, [SPOUSE [[PARENT1], [PARENT2]]] 400 13 08, 400 14 32, 400 15 27, 400 15 48; Henrietta,Mills,Hyde [Charles,Smith,Shelton [[George,Shelton], [Betsy,Wooster]]]>

Attributed relations:

In our implementation, each extracted relation includes the unique ID, Page, Line, and Offset of every participating token.

Acknowledgment:

To my long- time friend, colleague and collaborator
Emeritus Professor Dave E. Embley of Brigham Young University, and his team

Merci!

Project initiated with and at the suggestion of
my long-time friend and collaborator
Professor Emeritus **David W. Embley** of Brigham Young University.

All experimental data, and much other help, from
Prof. Embley and his colleagues at **BYU** and **FamilySearch**.

C’est tout.