Historical life cycle reconstruction by indexing

Gerrit Bloothooft
Jelte van Boheemen
Marijn Schraagen

Utrecht institute of Linguistics, Utrecht University, The Netherlands

g.bloothooft@uu.nl; jeltevanboheemen@gmail.com; m.p.schraagen@uu.nl

application domain

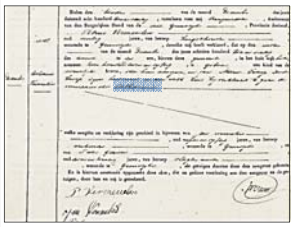
The Netherlands, province of Zeeland
1811 – 20th century
vital registration

The Zeeland challenge

certificates

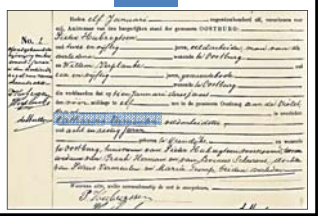
Catharina Vermeulen

birth



marriage

decease



a life cycle

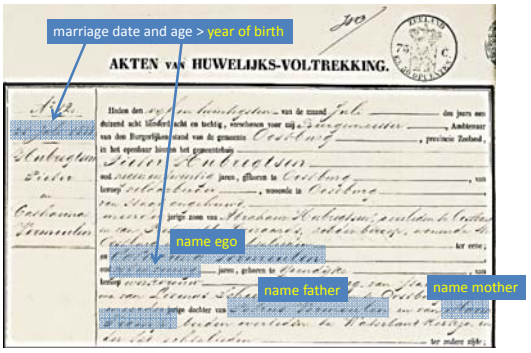
Catharina Vermeulen

1842: **Catharina Vermeulen is born**.
1872: Catharina Vermeulen, age 29,
marries Izaak Herman, age 32.
1872: child Maria Herman is born.
1874: child Tannetje Herman is born.
1875: child Tannetje Herman, age 0, dies.
1876: child Tannetje Herman is born.
1876: child Tannetje Herman, age 0, dies.
1877: child Catharina Herman is born.
1878: child Petrus Herman is born.
1881: partner Izaak Herman, age 41, dies.
1888: child Maria Herman, age 15, dies.
1900: child Petrus Herman, age 22,
marries Suzanna Beerens, age 20.
1905: child Catharina Herman, age 27,
marries Marinus Moes, age 23.
1939: child Petrus Herman, age 61, dies.

1881: Catharina Vermeulen, age 38,
marries Livinus Scheerens, age 52.
1882: child Abraham Scheerens is born.
1883: partner Livinus Scheerens, age 54, dies.
1903: child Abraham Scheerens, age 20,
marries Adriana Simpelaar, age 19.
1928: child Abraham Scheerens, age 45,
marries Catholijntje Cornelis, age 39.
1954: child Abraham Scheerens, age 71, dies.

1888: Catharina Vermeulen, age 45,
marries Pieter Hubregtsen, age 29.
1911: **Catharina Vermeulen**, age 68, **dies**.

certificate 3rd marriage



who is who: uniquely identifying information

- name ego, date and place of birth
- name ego, year of birth, names of parents
- name ego, year of birth, name of partner
- ....

factors determining uniqueness

- size of population
- data completeness
- data accuracy
- name conventions
- social/family relationships

Zeeland

population size (census)

year	people
1830	137,200
1869	177,569
1930	247,500

in-migration (census 1869)

born in	people	cumulative %
same municipality	118,137	66.0
Zeeland	46,016	92.4
The Netherlands	7,810	96.8
Belgium	5,213	99.8
other	393	100.0

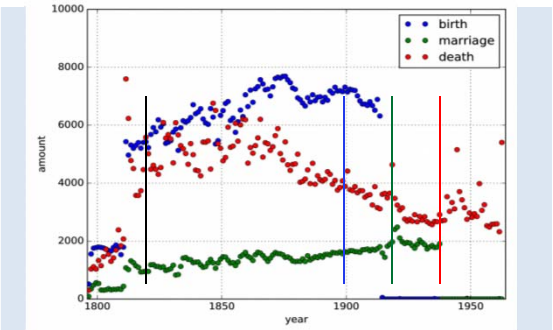
available data

1,558,205 certificates and 5.6 million individual tokens
(key information digitized with the help of volunteers since 1990)

type	certificates	people mentioned (million)	range
birth	698,285	2,1	1811-1913
marriage	192,231	1,2	1811-1938
divorce	1,690	-	1811-1938
death	665,999	2,3	1811-1963

LINKS Zeeland Cleaned Dataset (Marriages, Births and Deaths), release 2016_01

certificates per year



matching of records

strategies:

pairwise edit-distance
- certificate-based or ego-based
- various (edit-)distance measures
- problems for large edit-distances
- quadratic computation problem

sorted neighborhood
- ego-based
- requires standardized data
- matching within window (size=2) only
- fast

sorting example

ego surname	ego first name	sex	mother surname	mother first name	father surname	father first name	date birth (days)	role
doorn	wilhelmina	f	dishoek	petronella	doorn	abraham	657040	deceased
doorn	wilhelmina	f	dishoek	petronella	doorn	abraham	657103	bride
doorn	wilhelmina	f	doorn	jacoba	doorn	johannes	671562	child
doorn	wilhelmina	f	doorn	jacoba	doorn	johannes	671566	deceased
doorn	wilhelmina	f	nauta	geertrui	doorn	kornelis	647143	deceased
doorn	wilhelmina	f	schorer	adriana	doorn	hendrik	661241	child
doorn	wilhelmina	f	vlag	maria	doorn	leendert	697569	bride
doorn	wilhelmina	f	vlag	maria	doorn	leendert	697661	child

optimal field order for sorting key

for birth, marriage, death of ego

1. surname ego [distinguishes more than first name]
2. first name ego
3. sex ego
4. surname mother
5. first name mother
6. surname father [father not always known]
7. first name father
8. birth date ego [imprecise]

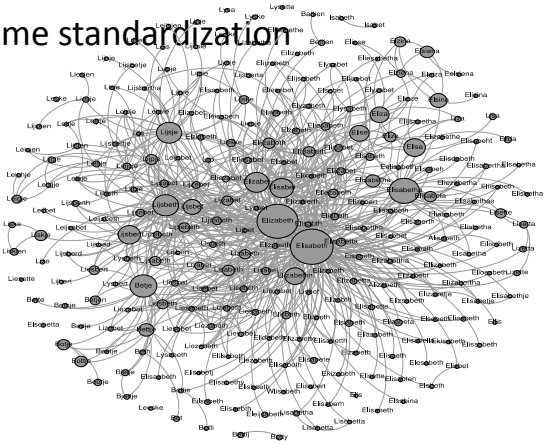
place of birth ego is not used [imprecise]

two sorting keys

Birth	ego mother father birth-date	
Marriage	ego mother father age	ego partner
Death	ego mother father age	ego (partner)
Children		ego partner
Death partner		(ego) partner

marriage is bridge between two sortings

name standardization



name standardization

first names

- 21,157 different initial first names (5.6 million tokens)
- 14,163 names with 762 standards [99.06% of tokens]
- 6,994 names not-standardized [0.94% of tokens]

surnames (without particles: *de Vries*)

- 51,380 different surnames
- 39,335 semi-phonetized
- 12,782 limited to first 4 characters

matching of subsequent records

- requirements
  - all *minus one* standardized names match
  - same sex
  - date difference < 400 days

this results in blocks of records
which internally subsequently match

results

- no golden standard
- consistency check (events chronologically correct)
  - 508,862 consistent blocks
  - 2,939 inconsistent blocks (0.56%)
- IISH Amsterdam: edit-distance on certificates
  - 31,773 extra links at IISH (missing parents, missing dates, standardization error)
  - 236,959 links missing (incomplete analysis infant mortality, spelling variation in 1-4 names, cleaning for duplicate records)

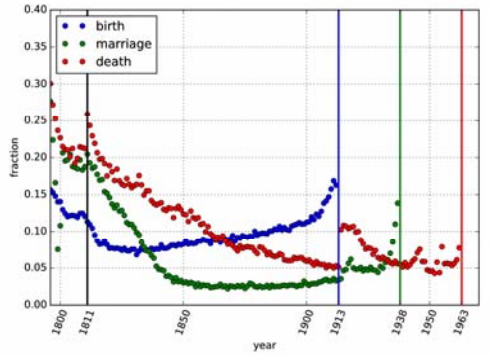
non-matching original fields

between links within consistent block

Levenshtein distance	ego surname	mother surname	ego first name	mother first name	father first name
0	94.0	91.3	88.3	88.1	89.6
1	5.0	6.3	7.2	6.8	5.3
2	0.8	1.6	2.0	1.8	1.7
>=3	0.2	0.8	2.5	3.3	3.4

every 10 links have on average at least one name with edit distance >=3

non-linked ego records



life courses

we have ego: birth, marriage, decease
now add children and integrate more partners

second sorting key (all match) :

1. surname ego
2. first name ego
3. sex ego
4. surname partner
5. first name partner

integrate results into previously obtained blocks (with rules)

+ children (birth, marriage, death)
+ partners (death)

life courses (per marriage)

- 285,583 consistent life course *sections*
  per partner of ego (with on average 11 events)

event	fraction
birth ego	0.28
marriage ego	1.00
death ego	0.63
birth child	3.44
marriage child	1.52
death child	2.39
death partner	0.62

- 20,061 life courses with multiple marriages

a life cycle

1842: **Catharina Vermeulen is born**.
1872: Catharina Vermeulen, age 29,
marries Izaak Herman, age 32.
1872: child Maria Herman is born.
1874: child Tannetje Herman is born.
1875: child Tannetje Herman, age 0, dies.
1876: child Tannetje Herman is born.
1876: child Tannetje Herman, age 0, dies.
1877: child Catharina Herman is born.
1878: child Petrus Herman is born.
1881: partner Izaak Herman, age 41, dies.
1888: child Maria Herman, age 15, dies.
1900: child Petrus Herman, age 22,
marries Suzanna Beerens, age 20.
1905: child Catharina Herman, age 27,
marries Marinus Moes, age 23.
1939: child Petrus Herman, age 61, dies.

1881: Catharina Vermeulen, age 38,
marries Livinus Scheerens, age 52.
1882: child Abraham Scheerens is born.
1883: partner Livinus Scheerens, age 54, dies.
1903: child Abraham Scheerens, age 20,
marries Adriana Simpelaar, age 19.
1928: child Abraham Scheerens, age 45,
marries Catholijntje Cornelis, age 39.
1954: child Abraham Scheerens, age 71, dies.

1888: Catharina Vermeulen, age 45,
marries Pieter Hubregtsen, age 29.
1911: **Catharina Vermeulen**, age 68, **dies**.

jigsaw of life courses

test life courses that interact

- for additional information

1888: Catharina Vermeulen, age 45,
marries **Pieter Hubregtsen**, age 29.
1911: Catharina Vermeulen, age 68, dies.

1928: Pieter Hubregtsen, age 69, dies.
no mention of Catharina Vermeulen, because of 2nd marriage

- for consistency
  - marriage sequence (death of earlier partner)
  - child sequence (death of older child with same name)

issues for discussion

- need for standardization
  - level of standardization, multiple standards
- sorting under conditions of less rich information
  - application domains
- use of patronymics in sorting
  - missing surnames (<1811)
- step-wise building of life cycles
  - Bertrand Russell: whenever possible, substitute constructions out of known entities for inferences to unknown entities
- golden standard and comparison procedures
  - testing the quality of results
  - Occam's razor: the least number of individuals that can explain all data

to be continued on Thursday

The Zeeland challenge