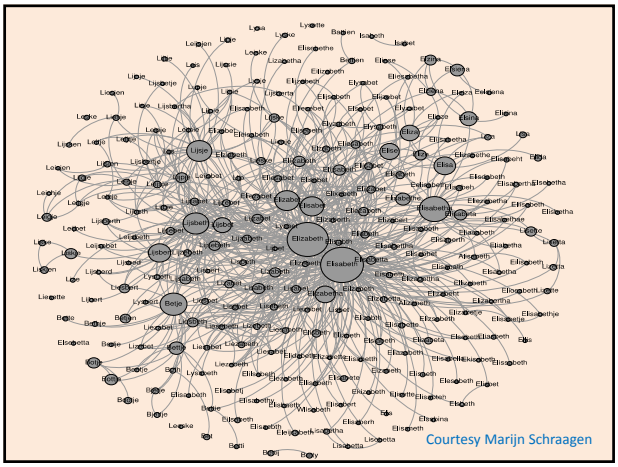
NAMES

Gerrit Bloothooft & David Onland, UiL-OTS Utrecht
Martin Reynaert, TiCC / Tilburg University
Katrien Depuydt & Tanneke Schoonheim, INT Leiden

aims

standardization of 19th century person names for
richer search results
OCR post-processing
nominal record linkage
onomastic research

resulting NAMES corpus
in RDF format for Linked Open Data
& lexicon service



Courtesy Marijn Schraagen

person name standardization

main issues

- on what basis?
  - onomastic expert knowledge
  - automatic learning from data
- data structure
  - handling ambiguities
- usage
  - search tools
  - nominal record linkage

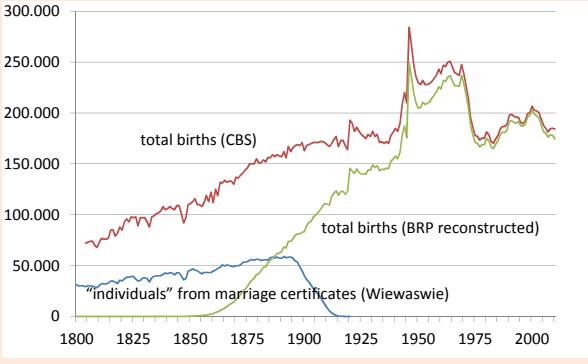
automatic learning from data

proven name variants

from *identified* individuals in various documents

followed by expert review

population and resources



base material

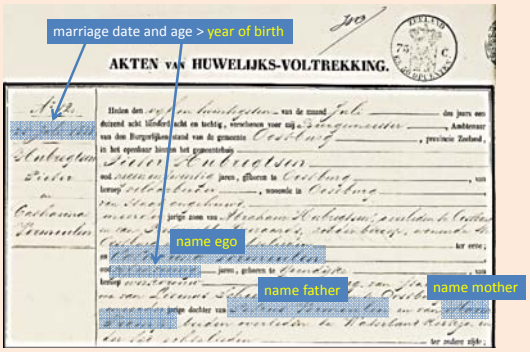
564.000 surnames and 189.000 first names

from 19th century sources
(birth, marriage, death certificates)

with 63 million person tokens

from Catch LINKS project – Wiewaswie (2009 version)

certificate 3rd marriage



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 (geographic spread)
- data completeness
- data accuracy
- name conventions
- social/family relationships

sufficient information

- first name ego
- three out of four names of parents

- one name may show up as variant
  (first name or surname)
  and variants can be harvested

proven-variant clustering

- surnames 127.154 into 15.114 standards

  compare:
  - corpus of Dutch surnames (CBG)
    320.000 names, 59.000 in NAMES, 11.800 standards
  - surnames in Belgium and northern France (Debrabandere),
    125.000 names, 49.000 in NAMES, 19.000 standards

- first names 42.196 into 2.900 standards
  (gender-dependent)

  compare:
  - first name dictionary (van der Schaar)
    20.000 names, 12.500 in NAMES, 2.400 standards

Christoffel

CCHRISTOFFEL
CGRISTOFFEL
CHIRSTOFFEL
CHISTOFFEL
CHISTOFFER
CHISTOPH
CHRESTOFF
CHRIFFEL
CHRIJSSTOPHORUS
CHRIJSTOPHORUS
CHRIOSTOFFEL
CHRISOFFEL
CHRISSTOFFEL
CHRISSTOFFELINA
CHRISSTOFFER
CHRISSTOFH
CHRISTAPHORUS
CHRISTEL
CHRISTEPH
CHRISTHOFFEL
CHRISTHOP
CHRISTHOPH
CHRISTHOPHORUS
CHRISTIOFFEL
CHRISTIPHORUS
CHRISTOBH
CHRISTOF
CHRISTOFEL
CHRISTOFELINA
CHRISTOFENA
CHRISTOFER
CHRISTOFF
CHRISTOFFE
CHRISTOFFEL
CHRISTOFFELINA
CHRISTOFFELL
CHRISTOFFELLINA
CHRISTOFFER
CHRISTOFFES
CHRISTOFFLE
CHRISTOFHE
CHRISTOFINA
CHRISTOFLE
CHRISTOFORA
CHRISTOFORIS
CHRISTOFORUS
CHRISTOFUS
CHRISTOHORA
CHRISTOKE
CHRISTOOFFEL
CHRISTOP
CHRISTOPF
CHRISTOPFEL
CHRISTOPH
CHRISTOPHARUS
CHRISTOPHE
CHRISTOPHEL
CHRISTOPHELT
CHRISTOPHER
CHRISTOPHERUS
CHRISTOPHIENA
CHRISTOPHINA
CHRISTOPHLE
CHRISTOPHOLUS
CHRISTOPHONES
CHRISTOPHONIS
CHRISTOPHONUS
CHRISTOPHORA
CHRISTOPHORE
CHRISTOPHORES
CHRISTOPHORI
CHRISTOPHORIS
CHRISTOPHORUS
CHRISTOPHOSUS
CHRISTOPHOZA
CHRISTOPHOZUS
CHRISTOPHRE
CHRISTOPHURUS
CHRISTOPHURUS
CHRISTOPHUS
CHRISTOPORUS
CHRISTOTINA
CHRISTPH
CHRISTPHEL
CHRISTPHORIS
CHRISTPHORUS
CHRISTSTOFFEL
CHRISTSTOFFELINA
CHRISTSTOFFER
CHRISTTOFFEL
CHRISTTOPH
CHRITOFFEL
CHROTOPEL
CHRSITOFFEL
CHTISTOFFEL
CRISSTOFFEL
CRISSTOFFER
CRISTOF
CRISTOFFEL
CRISTOFFER
CRISTOFORUS
CRISTOPH
CRISTOPHE
CRISTOPHORUS
KIRSTOFFEL
KISTOFFEL
KRISTOFFEL
KRISTOFFER
STEFFERTIEN
STIFFEK
STIOFFEL
STIVELIENA
STOFER
STOFFEA
STOFFEL
STOFFEN
STOFFER
STOFFERT
STOFFLE
STOFVER
STOFVERS
STOKKER
STOPFFER
ZOFFER

Christiaans

CHIRSTIAANSE
CHISTIAANS
CHRIATIAANS
CHRIESTIAENS
CHRIESTJAENS
CHRISATIAANS
CHRISITAANS
CHRISJANS
CHRISTAANS
CHRISTAANSE
CHRISTAEN
CHRISTAENS
CHRISTAENSSENS
CHRISTANI
CHRISTEAANS
CHRISTIAAENS
CHRISTIAAN
CHRISTIAANS
CHRISTIAANSE
CHRISTIAANSEN
CHRISTIAANSENS
CHRISTIAANSSENS
CHRISTIAANSZEN
CHRISTIAEN
CHRISTIAENS
CHRISTIAENSEN
CHRISTIAENSSENS
CHRISTIAN
CHRISTIANA
CHRISTIANE
CHRISTIANS
CHRISTIANUS
CHRISTJAANS
CHRISTRIAANSEN
CHRITIANS
CHRSITIAENS
CHTISTIAENS
CRISTIAAN
CRISTIAANS
CRISTIAANSE
CRISTIAANSEN
CRISTIAENS
CRISTIANS
KRISTIAANS
KRISTIAAN
KRISTIAANS
KRISTIANUS
KRISTIJAANS
KRISTIONS
KRISTJAAN
KRISTJAANS

Luijk

L?CKEN
L?CKERS
L?KEN
LAAKENS
LAEKENS
LAKENS
LAKIUS
LALENS
LOKKES
LOUKENS
LUCA
LUCCA
LUCK
LUCKA
LUCKE
LUCKEN
LUCKENER
LUCKER
LUCKERS
LUCKERT
LUCKES
LUCKNER
LUCKS
LUCKX
LUDKEN
LUEKEN
LUEKES
LUEKS
LUIBEN
LUICKX
LUIHEN
LUIJBEN
LUIJCK
LUIJCKEN
LUIJCKIJ
LUIJCKS
LUIJCKSE
LUIJCKX
LUIJCX
LUIJEX
LUIJK
LUIJK,
LUIJKE
LUIJKEN
LUIJKER
LUIJKES
LUIJKS
LUIJKX
LUIJM
LUIJN
LUIK
LUIKE
LUIKEN
LUIKENS
LUIKER
LUIKES
LUIKS
LUIKT
LUIN
LUIX
LUK
LUKA
LUKE
LUKEN
LUKENS
LUKER
LUKES
LUKHIEN
LUKIEN
LUKJE
LUKKA
LUKKEN
LUKKER
LUKKERS
LUKKERT
LUKKES
LUKKIEN
LUKNER
LUKS
LUKTE
LUKTJE
LUTKEN
LUUKS
LUX
LUYCK
LUYCKX
LUYCX
LUYK
LUYKE
LUYKEN
LUYKS
LUYKX
LUYN
LÜCKE
LÜCKEN
LÜCKER
LÜCKERS
LÜKE
LÜKEN
LÜKS

expert review

- surnames 15.114 standards into 11.539 standards
- first names 2.900 standards into 926 standards (gender-*in*dependent)

926 standards into ~500 base forms (first syllable etymology)

standards and application

choice of standards depends on application

– wide search: broad standards
– narrow search: fine standards

=> data structure for standards at different levels

data structure issues

full analysis

- learning variant properties
  (TICLL, Text-Induced Corpus CLean up)
  – context-dependent edit distance
  – clustered edit operations
- standardization of remaining names
  127.000 -> 564.000 surnames
  46.000 -> 189.000 first names

nominal record linkage

nominal record linkage

strategies:

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

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

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]

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

conclusion

- standardization based on proven variants
- expert review needed
- many uncertainties (short names)
- flexible data structure required
- extention to patronyms

- extremely useful for many applications
- likely useful for older documents