NAMES:

a Dutch corpus of person name variants

Gerrit Bloothooft (UiL-OTS), David Onland (UiL-OTS),
Martin Reynaert (TiCC), Katrien Depuydt (INT),
Matthieu Fannee (INT), Jauco Noordzij (Huijgens ING)

Person names are often central for search in (historical) documents. But names come in many variants. This concerns spelling variation, errors and aliases in original texts, but can also result from errors in optical character reading or automatic handwriting recognition.

Small variations can be efficiently handled by edit distance measures. But large differences such as between *Grietje* and *Margaretha* require standardization.

This project aims to develop a gold standard for 189.000 first names and 564.000 surnames from 19th century sources (63 million tokens), resulting from the LINKS and IMPACT projects.

ISSUES:

a) The definition of standards and hierarchical levels of standardization (variants of *Albert* and variants of *Elbert* taken together at a higher level)
b) Ambiguity in standards (*Jo* for *Johannes* and *Jozef, Mina* for **mina*), and unresolvable names, to be associated to many standards
c) Generalizability of standards beyond the 19th century

STEP 1: Collect name variants which can be proven to concern the same person

sources:

a) Output of record linkage on 19th century birth-, marriage-, and death certificates *which vary in a single name* but have sufficient evidence to be true links. This set of variants concerns 48.000 first names and 127.000 surnames.
b) Library thesauri which contain name variants of authors, proven by experts.

STEP 2: Expert review of standards associated to the variant set, supported by TICLL (Text-Induced Corpus CLean up). For TICCL this is a statistical learning phase as well.

STEP 3: Classification of the remaining 146.000 first names and 437.000 surnames to these standards, to a large extend automatically using TICCL (on the basis of adapted parameters).

STEP 4: Output in RDF format for Linked Open Data.

SURNAMES FROM *JACOB (2007)*

Ceuppens, Cobben, Cobbin, Cober, Cobet, Cobie, Cobus, Cobussen, Coobs, Coops, Cop, Cöp, Copijn, Copini, Copius, Coppee, Coppen, Coppens, Coppes, Coppy, Coppin, Coppis, Coppus, Cops, Cup, Cuppé, Cuppen, Cuppens, Cuppes, Jaakke, Jacob, Jacobi, Jacobse, Jacobsen, Jacobson, Jacobus, Jacobusse, Jacquemijns, Jakobs, Jakobsen, Keppens, Keuben, Kobbe, Kobben, Köbben, Kober, Kobes, Kobesen, Kobessen, Kobossen, Kobs, Kobus, Kobussen, Koobs, Koop, Koopen, Koops, Koopsen, Koopsma, Kop, Köpcke, Kopee, Kopjes, Köpke, Kopp, Köpp, Koppe, Koppei, Köppel, Koppen, Köppen, Koppenens, Koppens, Koppers, Köppers, Koppes, Koppeij, Koppies, Koppy, Koppijn, Koppius, Kops, Kubbe, Kubben, Kubbenga, Kubbinga, Kubes, Kubin, Kubis, Kup, Küp, Kuppen, Küppens, Küpper, Kuppers, Kuup, Jacobs, Yaacoub, Yaacoubi, Yaakoub, Yaakoubi, Yaakov, Yacob, Yacobi, Yacobus, Yacoob, Yacoub, Yacouba, Yacoubi, Yacoubian, Yacoubou, Yacubi, Yacubu, Yako, Yakob, Yakobchuk, Yakobi, Yakobowitz, Yakoob, Yakop, Yakoub, Yakoubi, Yakoubou, Yakub, Yakubi, Yakubovich, Yakubu

The project is submitted to the Clariah Research Pilot Call, but will be in focus in the UiL-OTS regardless.

CLARIAH
Common Lab Research Infrastructure for the Arts and Humanities