Record Linkage for Entities in RIAD using Deep Neural Networks

Thomas Gottron
Thomas.Gottron@ecb.europa.eu

EUROPEAN CENTRAL BANK
EUROSYSTEM

RIAD
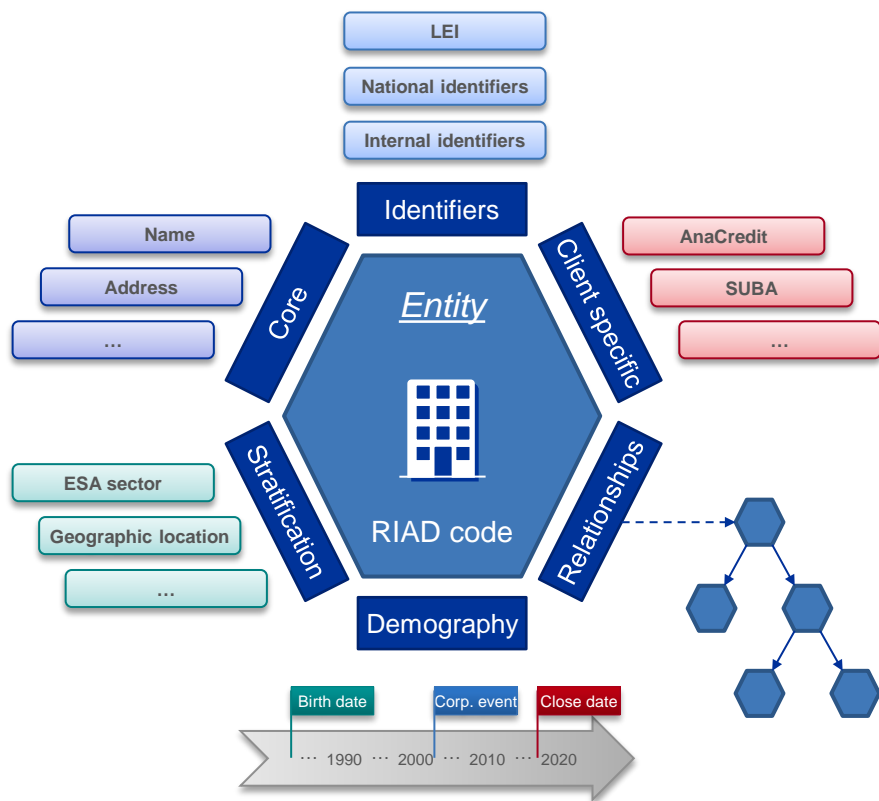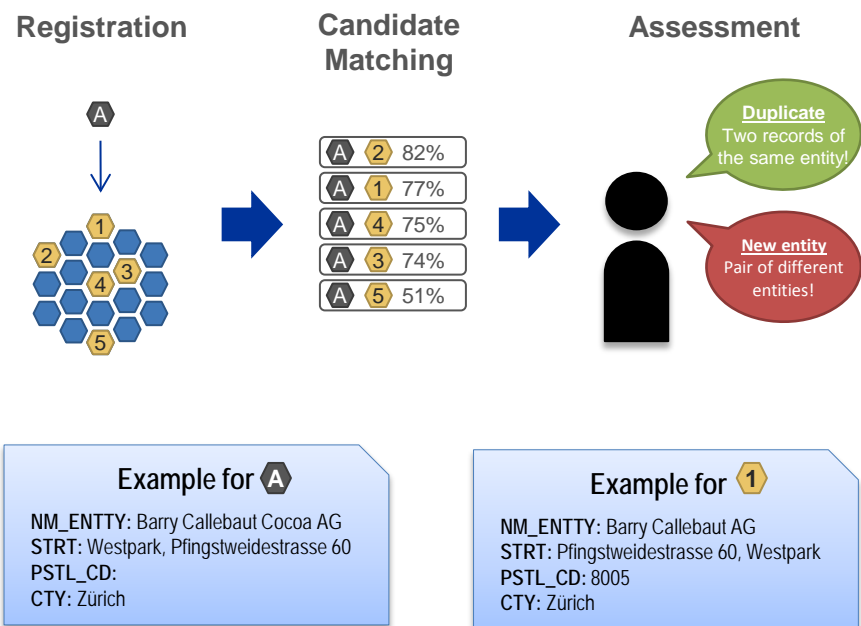*Register of Institutions and Affiliates Data*

RIAD

- ESCB's shared master dataset
- 9.7 million legal entities, incl. foreign branches
- Over 1,000 regular users across the ESCB and the SSM

Motivation

- Registration of non-resident entities (so called TEMP entities) requires assessment by host country
- Assessment for duplicates or new entities supported by automated candidate matching
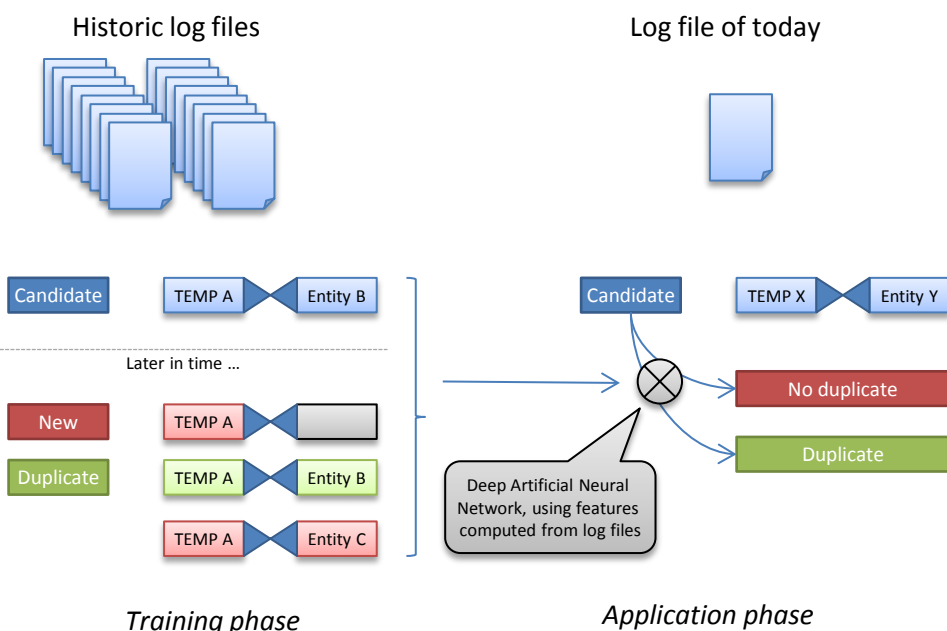- Final assessment is a manual process

Example for Ⓐ

NM_ENTTY: Barry Callebaut Cocoa AG
STRT: Westpark, Pfingstweidestrasse 60
PSTL_CD:
CTY: Zürich

Example for ①

NM_ENTTY: Barry Callebaut AG
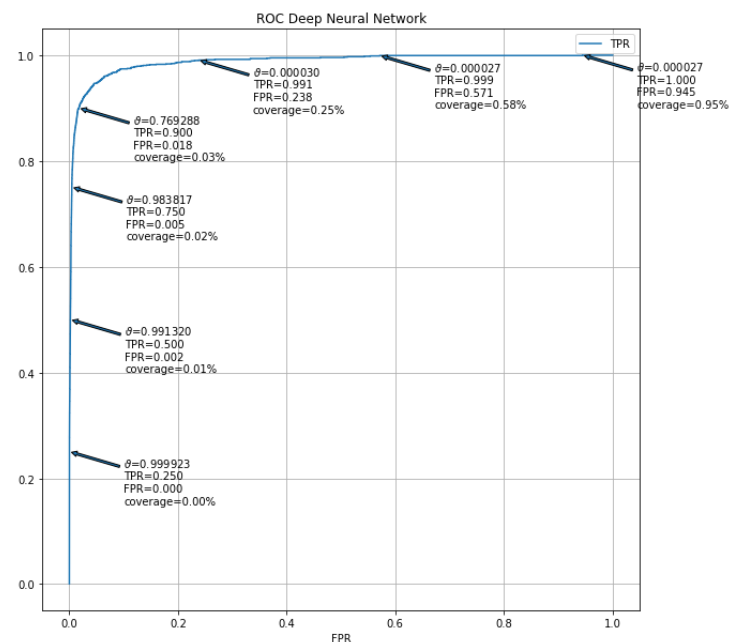STRT: Pfingstweidestrasse 60, Westpark
PSTL_CD: 8005
CTY: Zürich

Approach

- Log files provide information on candidates and results of manual assessments
- More than 1 year of observations
- 14,086 matching, 788,896 non-matching pairs

Training phase

Application phase

Evaluation

- Cross validation (80/10/10): 96.4% accuracy
- ROC curve for selecting confidence thresholds

ROC Deep Neural Network

θ=0.000030
TPR=0.991
FPR=0.238
coverage=0.25%

θ=0.000027
TPR=0.999
FPR=0.571
coverage=0.58%

θ=0.000027
TPR=1.000
FPR=0.945
coverage=0.95%

θ=0.769288
TPR=0.900
FPR=0.018
coverage=0.03%

θ=0.983817
TPR=0.750
FPR=0.005
coverage=0.02%

θ=0.991320
TPR=0.500
FPR=0.002
coverage=0.01%

θ=0.999923
TPR=0.250
FPR=0.000
coverage=0.00%

- Manual assessment on 2,485 entities in production

Accuracy	99.91%
Recall	99.69%
Precision	99.69%