KATHOLIEKE UNIVERSITEIT
LEUVEN

HARMONIZING PATENTEES

Julie Callaert – KU Leuven ECOOM

WIPO Name Harmonization Workshop
May 2019

AGENDA

- What we do and why we do it
- Harmonization approach
- Challenges & Opportunities
- Ways forward?

WHAT WE DO & WHY WE DO IT

KU Leuven - ECOOM – Technometrics

EEE-PPAT

- Applicant name harmonization (>> PSN_name)
- Inventor name disambiguation
- Applicant sector allocation

OTHER ENHANCEMENTS

- Regionalization of inventor and applicant addresses
- Characterization of non-patent references; matching of NPRs to Web of Science
- Domain concordance schemes (between science, technology and business)
- Matching between applicants and business repositories

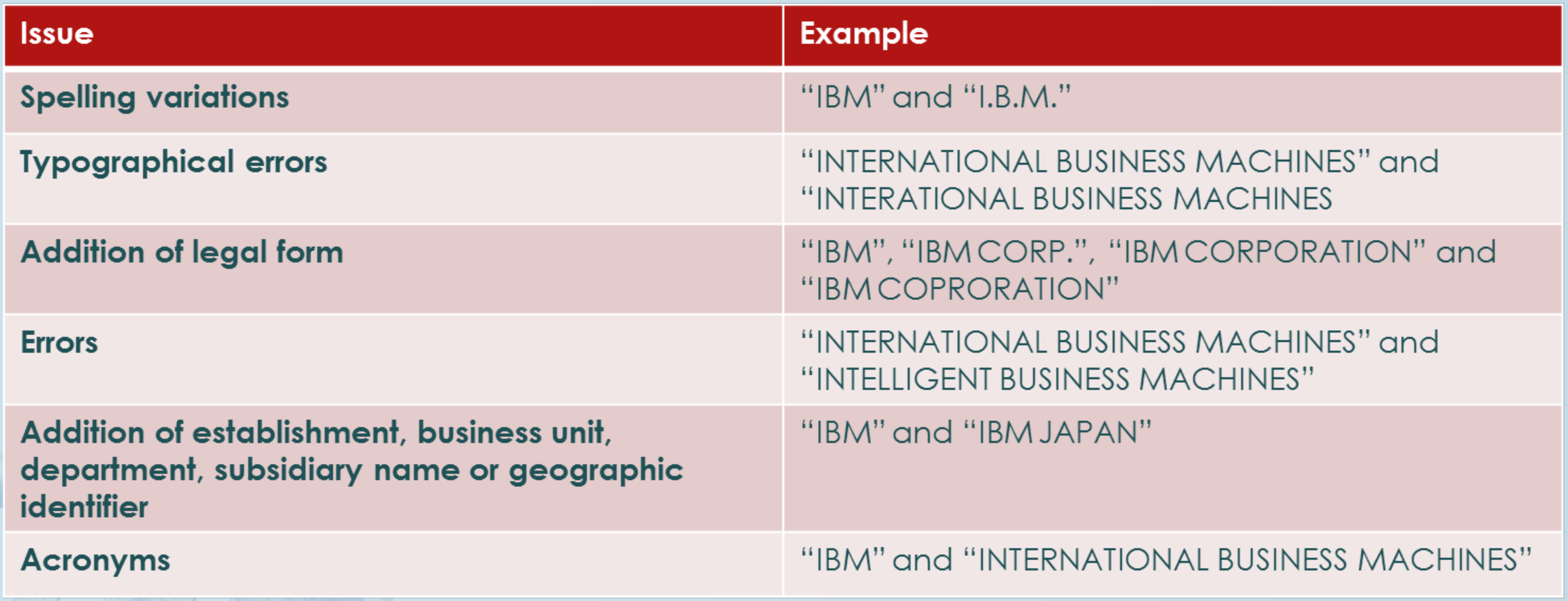
WORK IN PROGRESS

- Gender tagging of inventors
- Consolidation of applicants
- Text mining algorithms / machine learning

Why the need to harmonize?

- Applicant and inventor names in patent databases: idiosyncratic inputs
- No standardized format
- Use of different name variants within and across databases
- Spelling variations, typos, legal form addition, abbreviations, etc.
- E.g. 658 name variants (~ 1.068 PERSON_IDs) of “I.B.M”; 488 name variants (~ 1.491 PERSON_IDs) of “PANASONIC CORPORATION”

Why the need to harmonize?



Issue	Example
Spelling variations	“IBM” and “I.B.M.”
Typographical errors	“INTERNATIONAL BUSINESS MACHINES” and “INTERATIONAL BUSINESS MACHINES
Addition of legal form	“IBM”, “IBM CORP.”, “IBM CORPORATION” and “IBM COPRORATION”
Errors	“INTERNATIONAL BUSINESS MACHINES” and “INTELLIGENT BUSINESS MACHINES”
Addition of establishment, business unit, department, subsidiary name or geographic identifier	“IBM” and “IBM JAPAN”
Acronyms	“IBM” and “INTERNATIONAL BUSINESS MACHINES”

Why the need to harmonize?

- Listing and counting patents from 1 organization requires taking into account all name variants

- Failure to do so:
  - severe underestimation of an entity's patent portfolio
  - impedes name-based matching between patent databases and other information sources (like business registries or bibliographic databases)

- The extraction of reliable indicators is contingent on extensive efforts in data cleaning and enhancement

NAME HARMONIZATION APPROACH

Applicant name harmonization

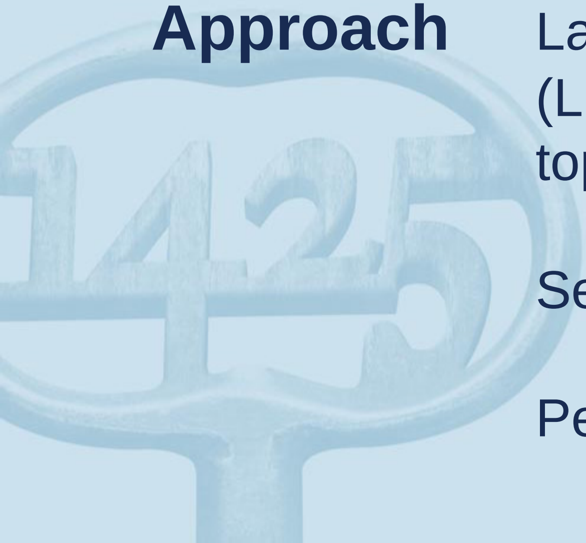
Target ‘person’-table in PATSTAT (~ 16.000.000 names)

Objective to harmonize different name variations occurring for one and the same applicant (→ PSN_name)

Approach Layered: combination of fully automated procedure (L1; all applicants) and further “manual” cleaning of top applicants (N = 2700) to increase recall (L2)

Self-referential

Performed upon each PATSTAT release

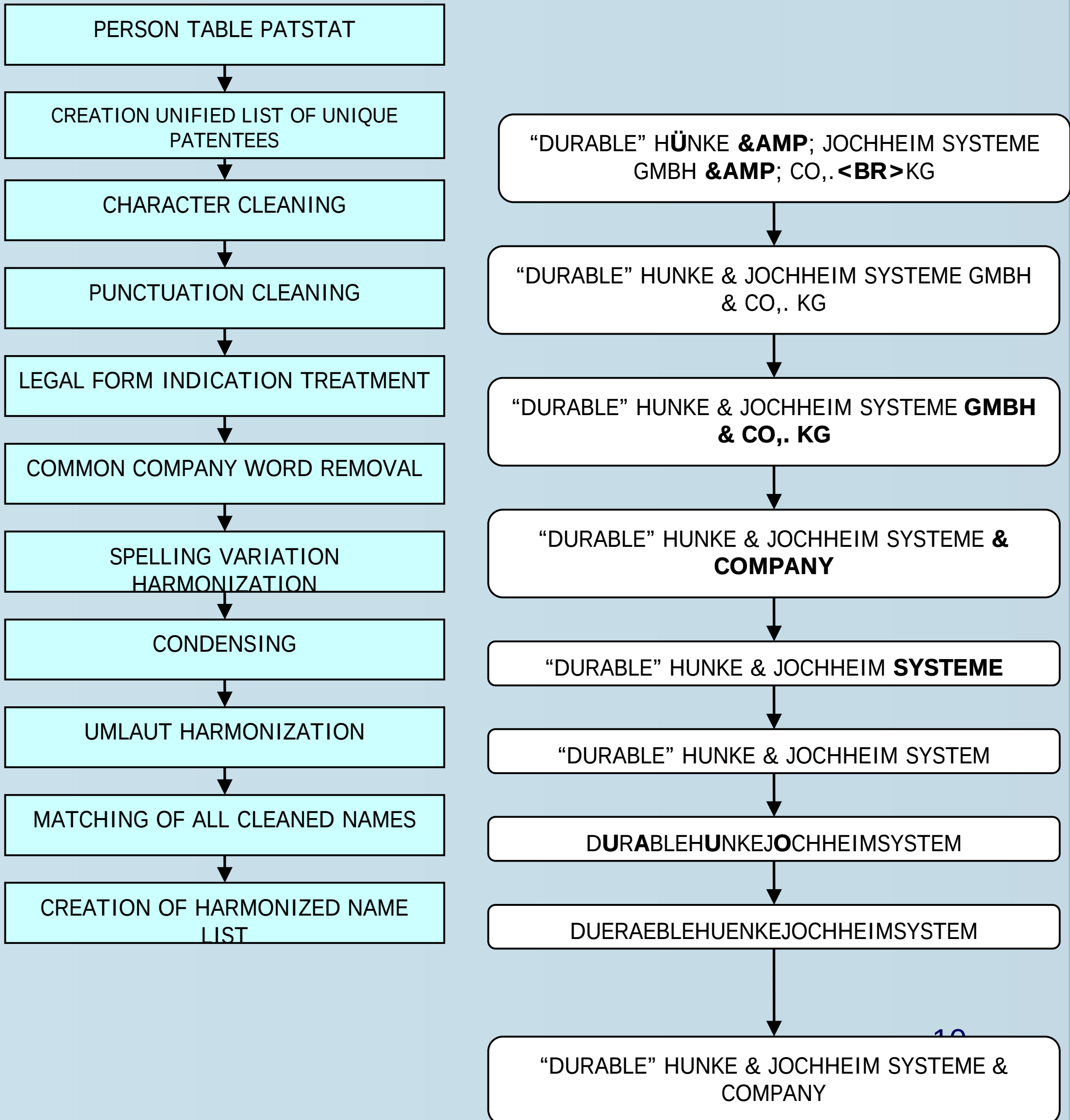


Applicant name harmonization

Layer 1: Automated procedure

Results:

- 21% reduction of unique names (from 15.969.238 to 12.547.700 names)
- 27% increase in patent volume per applicant
- > 99% accuracy



PERSON TABLE PATSTAT

CREATION UNIFIED LIST OF UNIQUE PATENTEES

CHARACTER CLEANING

PUNCTUATION CLEANING

LEGAL FORM INDICATION TREATMENT

COMMON COMPANY WORD REMOVAL

SPELLING VARIATION HARMONIZATION

CONDENSING

UMLAUT HARMONIZATION

MATCHING OF ALL CLEANED NAMES

CREATION OF HARMONIZED NAME LIST

“DURABLE” H**Ü**NKE **&AMP**; JOCHHEIM SYSTEME GMBH **&AMP**; CO,.**<BR>**KG

“DURABLE” HUNKE & JOCHHEIM SYSTEME GMBH & CO,. KG

“DURABLE” HUNKE & JOCHHEIM SYSTEME **GMBH & CO,. KG**

“DURABLE” HUNKE & JOCHHEIM SYSTEME **& COMPANY**

“DURABLE” HUNKE & JOCHHEIM **SYSTEME**

“DURABLE” HUNKE & JOCHHEIM SYSTEM

D**U**R**A**BLEH**U**NKEJ**O**CHHEIMSYSTEM

DUERAEBLEHUENKEJOCHHEIMSYSTEM

“DURABLE” HUNKE & JOCHHEIM SYSTEME & COMPANY

Applicant name harmonization

Layer 2: Further ‘manual’ cleaning of top applicants

- Complementary step - improving recall
  - Starting point: harmonized applicant names resulting from previous layer 1
  - Selection of top applicants (by technological field): > 2700 names treated
  - Approximate string searching on condensed names, using ‘Levenshtein distance’
  - Validation (human rating) of suggested matches
  - Accounting as well for name changes (harmonization to most recent name)

	After Level 2 harmonzation (HRM_L2)	**After Level 1 harmonization (HRM_L1)**	**Original PATSTAT name**
Nbr of distinct Names	2726 (99% red.)	100280 (52% red.)	207955
Avg Nbr of matched patents per name	11733,41 (x76)	318,96 (x2)	153,81

Applicant name harmonization

Result: shifts in ranking after name harmonization

Rank	Original name	Patent count
1	SAMSUNG ELECTRONICS CO., LTD.	412246
2	MATSUSHITA ELECTRIC IND CO LTD	354095
3	HITACHI LTD	334561
4	TOSHIBA CORP	299889
5	CANON INC	279936
6	IBM	251777
7	MITSUBISHI ELECTRIC CORP	250804
8	NEC CORP	226190
9	LG ELECTRONICS INC.	213256
10	ROBERT BOSCH GMBH	210477
11	FUJITSU LTD	203388
12	GENERAL ELECTRIC COMPANY	192791
13	SIEMENS AKTIENGESELLSCHAFT	179473
14	SONY CORP	174772
15	RICOH CO LTD	161786

Rank	Harmonized name	Patent count
1	PANASONIC CORPORATION	676975
2	TOSHIBA CORPORATION	589202
3	HITACHI	527728
4	SAMSUNG ELECTRONICS COMPANY	520447
5	CANON	501468
6	MITSUBISHI ELECTRIC CORPORATION	431759
7	NEC CORPORATION	427654
8	SONY CORPORATION	377221
9	SIEMENS	374577
10	FUJITSU	371484
11	IBM	346367
12	PHILIPS ELECTRONICS	322230
13	TOYOTA MOTOR CORPORATION	291123
14	GE (GENERAL ELECTRIC COMPANY)	284576
15	ROBERT BOSCH	276823

CHALLENGES & OPPORTUNITIES

Challenges & Opportunities

- **Applicant harmonization:**
  - Our chosen L2 approach implies that for one company (group), there may still be distinct level_2 harmonized names.

hrm_l2	Patent count
SERVICES PETROLIERS SCHLUMBERGER	2012
SCHLUMBERGER TECHNOLOGY	1964
SCHLUMBERGER HOLDINGS	1716
SCHLUMBERGER	347
SCHLUMBERGER INDUSTRIES	290

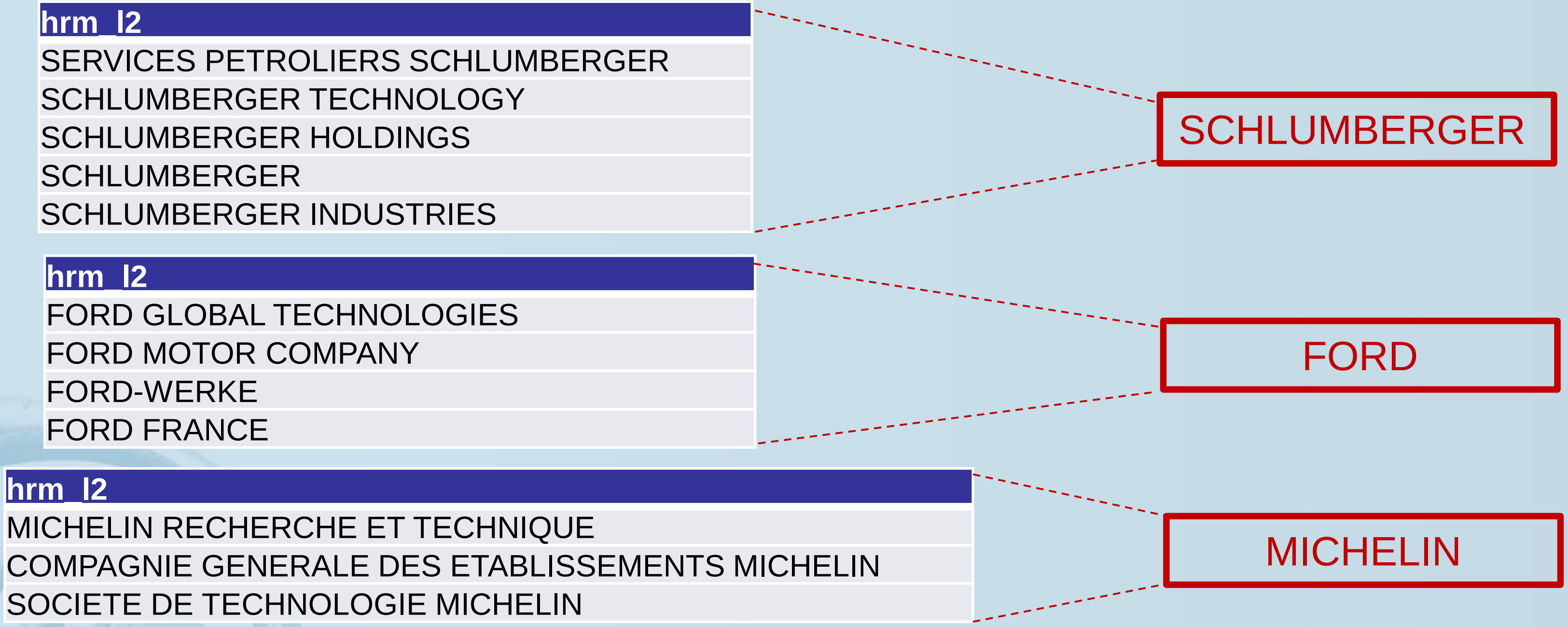
hrm_l2	Patent count
FORD GLOBAL TECHNOLOGIES	2150
FORD MOTOR COMPANY	2093
FORD-WERKE	1624
FORD FRANCE	1502

hrm_l2	Patent count
MICHELIN RECHERCHE ET TECHNIQUE	2164
COMPAGNIE GENERALE DES ETABLISSEMENTS MICHELIN	1971
SOCIETE DE TECHNOLOGIE MICHELIN	870

  - This is a feature, not a bug…

Challenges & Opportunities

- **Applicant harmonization:**
  - Our chosen L2 approach implies that for one company (group), there may still be distinct level 2 harmonized names.



hrm_l2
SERVICES PETROLIERS SCHLUMBERGER
SCHLUMBERGER TECHNOLOGY
SCHLUMBERGER HOLDINGS
SCHLUMBERGER
SCHLUMBERGER INDUSTRIES

SCHLUMBERGER

hrm_l2
FORD GLOBAL TECHNOLOGIES
FORD MOTOR COMPANY
FORD-WERKE
FORD FRANCE

FORD

hrm_l2
MICHELIN RECHERCHE ET TECHNIQUE
COMPAGNIE GENERALE DES ETABLISSEMENTS MICHELIN
SOCIETE DE TECHNOLOGIE MICHELIN

MICHELIN

  - This is a feature, not a bug…
  - **Adding a third layer?**

Challenges & Opportunities

- **Applicant harmonization:**
  - Consolidation
    - Reliable and up-to-date databases on M&A required
    - Complexities related to following up on passed trajectories of M&A, cases of subsequent demergers,…
    - M&A do not necessarily imply complete transfer of patent portfolios

WAYS FORWARD?

Ways forward?

A priori standardization

Eliminating the need for post-hoc harmonization

- Enforcing standardized input format for applicant / inventor names
- Applicant / Inventor ID-numbers that are assigned upon first patent application and that are to be inputted upon each new application
- Cross-datasource identifiers (VAT numbers,…)

Ways forward?

A posteriori treatment

Facilitating post-hoc harmonization

- Text mining applications: mapping topical architectures and topic overlap in patent portfolios on the level of individuals / organizations
- Mapping clusters of inventors / applicants, based on topical maps or network analysis of inventors / applicants

Ways forward?

The earlier on in the process identifiers can be integrated (i.e. if they would be imputed already formally in the phase of the patent application), the more efficient the process.

A posteriori identifiers: challenge remains to coordinate and make sure identifiers are consistent among databases (between editions of 1 database and between different databases).

Thank you

julie.callaert@kuleuven.be