WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Name Standardization Workshop 2019

CWS/7 item 8(c)

General

- Approved by CWS/6 in Oct 2018

- Held at WIPO headquarters May 2 & 3, 2019

- 30 participants representing 8 IP Offices and 9 organizations (data providers, researchers, applicants, user groups)

- Day 1 – Name Standardization topics

- Day 2 – Identifier topics

- Task Force meeting at end of each day

Name Standardization: Current Practices by IPOs

- Four IPOs (KIPO, EPO, IP Australia, USPTO) presented their practices

- All Offices reported issues with applicant data duplication

- Some Offices are working on data cleaning projects for consistency (punctuation, spacing, abbreviations, etc)

- USPTO economists have moved away from names to machine learning on other data points (location, title, etc) for more reliable owner / inventor determinations

- EPO use name normalization and unique identifiers for PATSTAT data

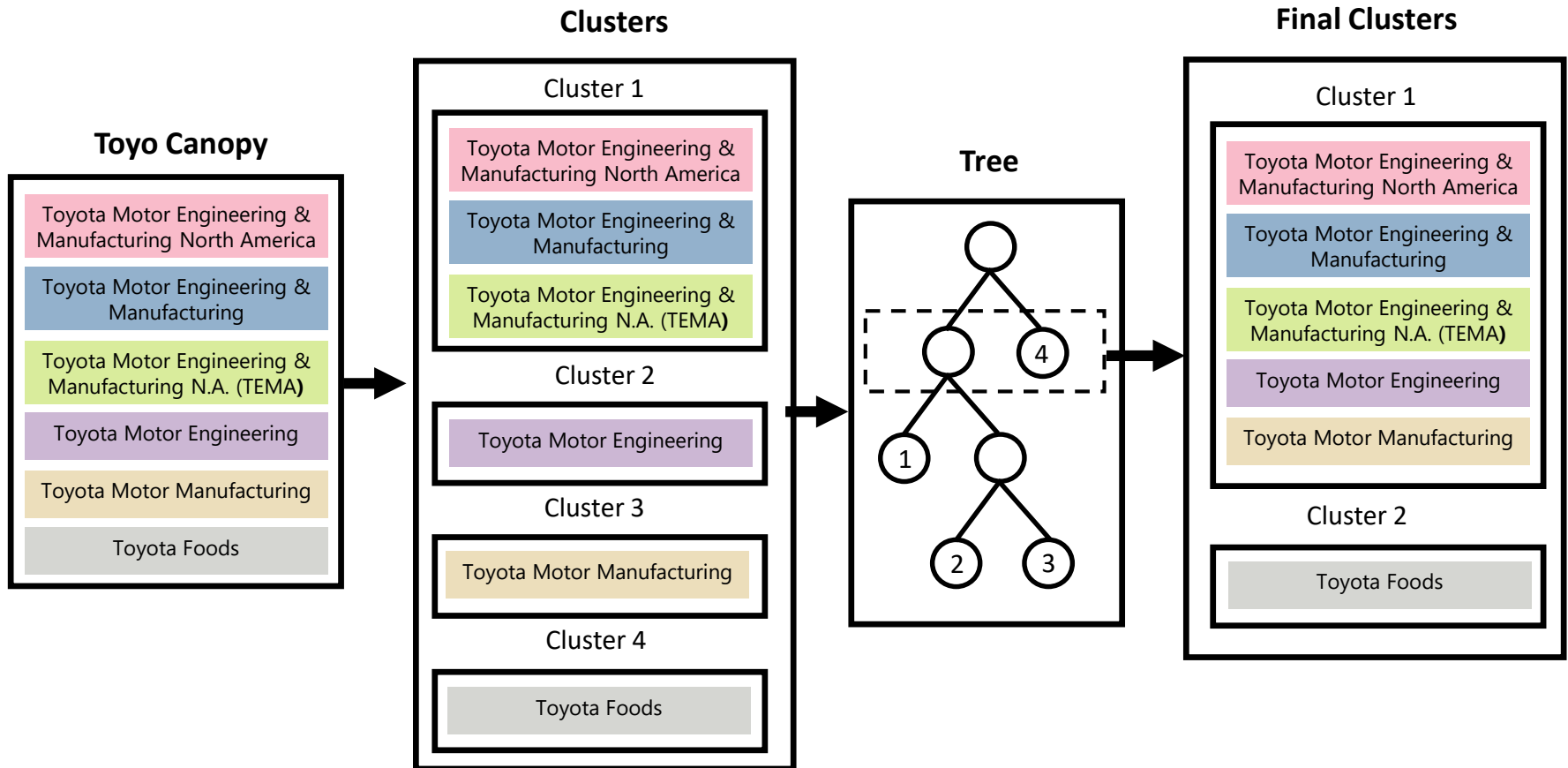# Phase 1 – Name & Address Clean-up – Merge Candidates

Merge Candidates:
1.19 M

- Full Stops

- Commas

- Ampersands

- Word CASE

- Word SPACING

Merge Candidates:
363,653

Agglomerative clustering for assignees

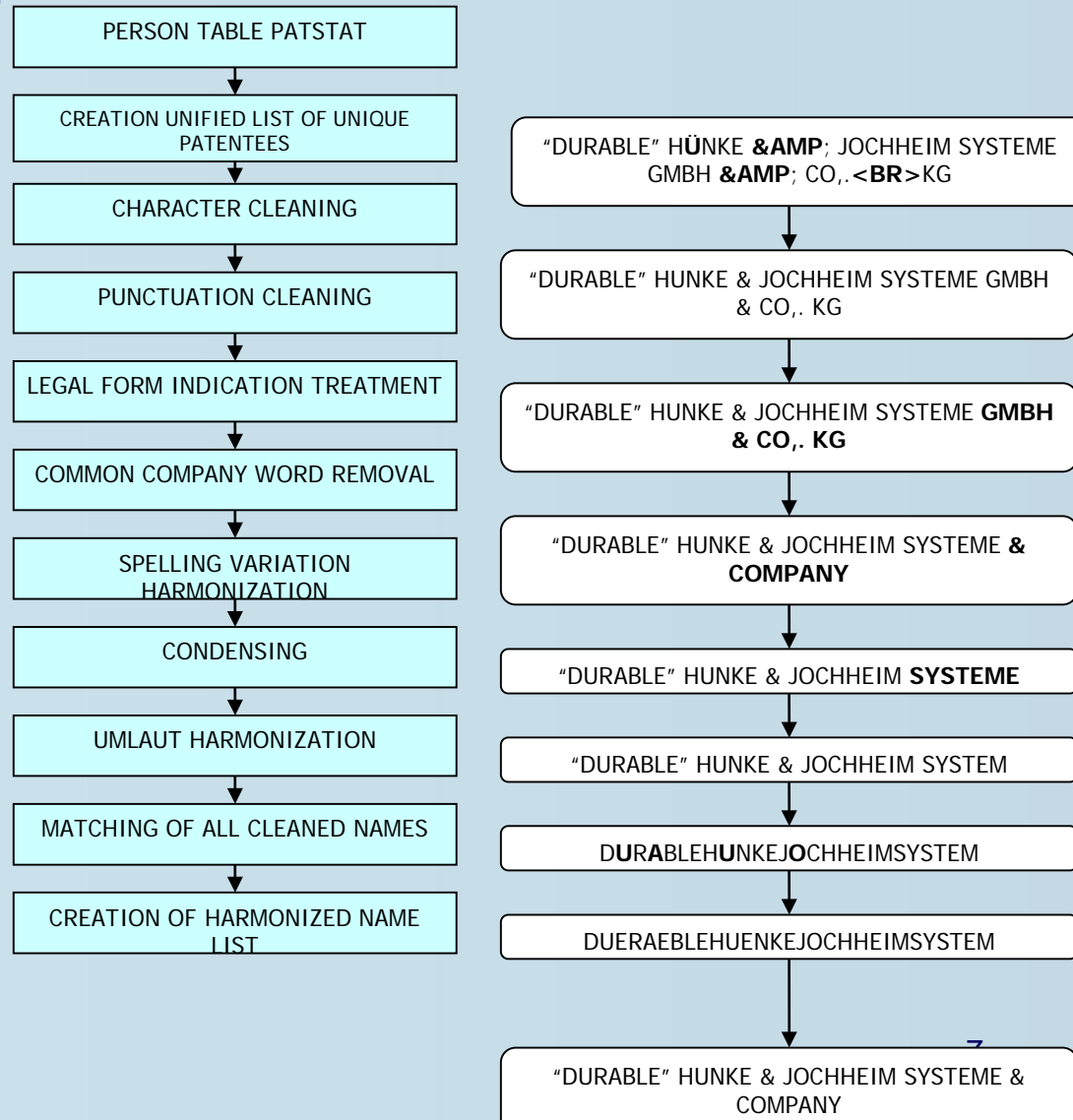# Name Standardization: Current Practices by Others

- KU Leuven performs automated name cleaning for EPO PATSTAT data, followed by manual name matching for top 2700 applicants

- WIPO ATAC is developing a machine learning tool for automated name transliteration across languages

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" HÜNKE **&AMP**; JOCHHEIM SYSTEME GMBH **&AMP**; CO,.**
**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**URA**BLEH**U**NKEJ**O**CHHEIMSYSTEM
↓
DUERAEBLEHUENKEJOCHHEIMSYSTEM
↓
"DURABLE" HUNKE & JOCHHEIM SYSTEME & COMPANY

Name Standardization: Roundtable discussion 1

- Industry increasingly relies on IP data for business decisions: landscaping, transactions, licensing, etc

- Analysis depends on the quality of data collected by IPOs

- Good data collection requires resources and procedures to be effective

- Many IPOs are limited in what data they must accept from applicants and what corrections they can make

- Changing data collection practices would require significant legislative and IT changes for many IPOs

Name Standardization: Roundtable discussion 2

- Two distinct uses of IP data emerged:

    1. Data collection / maintenance at IPOs for legal records

    2. Aggregate uses of IP data for statistics or analysis

- Legal records require high accuracy; even simple changes may introduce errors

- Aggregate uses can better tolerate minor errors for better overall results

Name Standardization: Roundtable discussion 3

- Sharing algorithms or code for name standardization may not be feasible at this time

- Different uses for IP data have different requirements for how names are cleaned and combined

- Technology in this field is evolving rapidly

- Some IPOs may face restrictions on sharing data or algorithms

  - e.g. code owned by contractors, privacy statutes, etc

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Name Standardization: Applicant Views

- Applicants reported that recorded name data can impact legal rights and obligations; even minor differences can be significant

  - e.g. ABC Corporation and ABC Incorporated

- Applicants may not mind minor corrections that do not affect legal rights

- Applicants would like the chance to review before any name corrections are made and reject them if needed

- Large filers would like to be notified of name corrections so they can update their internal systems

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Name Standardization: Conclusions

- There are significant differences in the data systems and legal requirements for the data each IPO stores

- Standardizing data collection or correction procedures across many IPOs is probably not feasible

- Standardizing formats for data sharing between offices is more realistic, but does not address the issues with name matching.

Identifiers:
Examples from Private Sector

- ISNI is an ISO standard for identifiers in the creative industries

- ISNI uses global, persistent identifiers with public lookups: http://www.isni.org/isni/000000012281955X

- Private sector registrars maintain names, based on public information (publishing and music industries)

- ORCID provides ISNI-compatible IDs for academic researchers, with more control over what data is publicly available

What does it look like?

for display
purposes only

15 decimal (base 10)
digits for persona

check digit
may be X

ISNI 0000 0004 2756 6266

- also http://isni.org/isni/0000000427566266
- public search interface, richer API for members, registration agencies
- numbering scheme shared with ORCID

IDENTIFIERS FOR PEOPLE, PLACES, AND THINGS

ORCID iD	First/given name	Last/family name	Other names	Affiliations
https://orcid.org/0000-0002-1701-6370	Matt	Buys		Scion, South African National Biodiversity Institute, North-West University, University of Stellenbosch Faculty of Science
https://orcid.org/0000-0001-7234-3684	Matt	Buys	Matthew Buys, MJ Buys, M Buys, Matthew J Buys	ORCID, University of the Witwatersrand, Wits Business School

Identifiers: Examples from IPOs

- 3 IPOs (IP Australia, UKIPO, Rospatent) presented

- Some offices moving to customer-based data models, which requires cleaning up current data

- Existing national identifiers (tax number, passport, etc) are used but identifiers for foreign applicants is difficult

- Challenges with resources, contractors, and incorrect data entry by applicants limits the effectiveness of identifiers

Identifiers: application form

title list app form:

Applicant Identifiers

ИДЕНТИФИКАТОРЫ ЗАЯВИТЕЛЯ

ОГРН — Main State Registration Number

КПП — Tax Reason Code

ИНН — Taxpayer Identification Number

СНИЛС — Insurance Individual Account Number

ДОКУМЕНТ (серия, номер) — Passport

КОД СТРАНЫ (если он установлен) — Country Code

before 2010 only "Main State Registration Number"

only internal use, mainly for tax considerations

Cleaning → Categorisation → Matching

ADP	CleanName	Company name	Company number
5593900	CROSS MANUFACTURING COMPANY (1938)	CROSS MANUFACTURING COMPANY (1938) LIMITED	342798
7528847	CROSS MANUFACTURING (1938) COMPANY LIMITED	CROSS MANUFACTURING COMPANY (1938) LIMITED	342798
392571	CROSS MANUFACTURING COMPANY (1938) LIMITED	CROSS MANUFACTURING COMPANY (1938) LIMITED	342798
4723706	CROSS MANUFACTURING COMPANY (1938) LIMITED	CROSS MANUFACTURING COMPANY (1938) LIMITED	342798
85369049	CROSS MANUFACTURING COMPANY (1938) LIMITED	CROSS MANUFACTURING COMPANY (1938) LIMITED	342798

Identifiers:
Roundtable discussion 1

- Quality of applicant name data has improved, but still not sufficient for locating all relevant IP documents

- Industry concerns over identifiers being used to link separate legal entities --> impact on legal rights

- For identifiers to be useful, IPOs must allocate resources and develop procedures for their use

- ORCID and ISNI stressed that identifiers are only one part of a solution to the problems faced by IPOs

- Largest cost of using identifiers may not be IT systems, but governance procedures

Identifiers:
Roundtable discussion 2

- One type of identifier may not cover all use cases; users need different granularity for different tasks
    - Ex: Toyota Motors Corp vs Toyota Motors Ltd vs Toyota Manufacturing

- Collecting and maintaining high quality applicant data by IPOs may be more useful to analysts than identifiers

- Some IPOs reported that introducing new fields in their data systems can be difficult or impossible

- Legal requirements can limit the ability of some IPOs to assign or request identifiers from applicants

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Identifiers:
Roundtable discussion 3

■ Participants agreed that applicants should be involved in any process to develop identifiers

■ IPOs might provide guidance for applicants on how to effectively manage their identifiers or other data

■ However, voluntary schemes may have low compliance

Task Force meetings

- Task Force discussed many options on name standardization and identifiers from the Workshop

- Changes to IPO data collection systems are not feasible for legal and technical reasons

- Improving quality of collected data is a common goal among members

- Due to different needs and environments at IPOs, does not make sense to develop software tools or algorithms at this time

- Developing global identifiers would be costly and the benefits are uncertain

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

Task Force – Next Steps

- Create forum for IPOs to share information and high level strategies for dealing with name collection and data cleanup issues

- Develop materials to raise awareness within IPOs of data quality issues

- Discuss the possibility of developing guidance for IPOs and applicants on data practices

■ Thank you for your attention!

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION