Name harmonization activities at the European Patent Office
Outline

• Customer data harmonization for the EP patent granting procedure
• Standardization of applicant names for DOCDB worldwide bibliographic database
• Harmonized names for the PATSTAT product line
Main difficulties → different forms of the name (address) to indicate the same applicant

- **Spelling variations**, e.g. “IBM” and “I.B.M.”, or “BAIN & CO” and “BAIN AND COMPANY”;

- **Typographical errors**, e.g. “INTERNATIONAL BUSINESS MACHINES” and “INTERATIONAL BUSINES MACHINES”;

- **Addition of the legal form** (again with possible acronyms, spelling variations, mistakes and typographical errors in the legal form) e.g. “IBM”, “IBM CORP”, “IBM CORPORATION” and “IBM CORPORATION”, or “BAYER”, “BAYER A.G.” and “BAYER AG”;

- **Errors** e.g. “INTERNATIONAL BUSINESS MACHINES” and “INTELLIGENT BUSINESS MACHINES”;

- **Other additions** like establishment, business unit, department, subsidiary name or geographic identifier, e.g. “IBM” and “IBM JAPAN”;

- **Use of acronyms**, e.g.”IBM” and “INTERNATIONAL BUSINESS MACHINES”.

---

EPO efforts in Customer data standardization

• Customer data relating to the Patent Granting procedure is maintained in a central Database containing many duplicates (ca. 20%)

• Duplicate customer data may arise from minimal differences in names or addresses

• The EPO will “clean” this database to create a central Master Data repository for EP applicants, with harmonized names and addresses (NACM Project)

• Aiming at single, unique identification per legal entity or individual
Quality at source!

• Cleaning strategy and plan is currently being developed to establish:
  1. Name and address standardization
  2. Deduplication of names

• Data Quality standards:
  – Addresses standardized to local Postal Authority databases
  – Names standardized based on syntactic analysis

• Data quality policies:
  – Data Steward role
  – Customer data process flows

• Challenges ahead:
  – Legal frameworks EPC, PCT
  – Validation of standardizations with applicants
DOCDB Standardization of Applicant Name - STAN

- **DOCDB** is the **EPO Worldwide Bibliographic Database**
- **STAN** since more than 30 years to improve patent search results

**Coverage:**
- For the **100 million patent documents** present in **DOCDB**:
  - +/- 250,000 standardized applicant names covering
  - +/- 945,000 variations
- Variations include typo/spelling errors, acronyms, etc.
- Standardization is only for **companies, universities, etc.** not for individuals

**Actual Challenges:**
- Achieve automated harmonization process of applicant names
- Maintain data quality using applicant name harmonization standards for the cleaning process
## STAN in Espacenet for the Public

### STAN names visible

<table>
<thead>
<tr>
<th>Inventor(s):</th>
<th>GILLESPIE JOHN JR [US]; LABEDZ RALPH H [US]; PLATT MICHAEL KENNETH [US]; SPANG RONALD H JR [US]; BERRILL JAMES FREI [US]; VOGEL MATTHEW STEPHEN [US]; GREANEY MICHELLE KOWALSKI [US] +</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicant(s):</td>
<td><strong>BAXTER</strong> INT INC [US] +</td>
</tr>
</tbody>
</table>

### STAN + nonstandardized names visible (clicking “+”)

<table>
<thead>
<tr>
<th>Inventor(s):</th>
<th>GILLESPIE JOHN JR [US]; LABEDZ RALPH H [US]; PLATT MICHAEL KENNETH [US]; SPANG RONALD H JR [US]; BERRILL JAMES FREI [US]; VOGEL MATTHEW STEPHEN [US]; GREANEY MICHELLE KOWALSKI [US] +</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applicant(s):</td>
<td><strong>BAXTER</strong> INT INC [US] + <em>(BAXTER INTERNATIONAL INC)</em></td>
</tr>
</tbody>
</table>
The EPO's PATSTAT product line

- Unique basis for conducting sophisticated statistical analyses of patent data
- Bibliographic data on nearly 100 million patent documents of leading industrialised and developing countries (DOCDB)
- Legal event data of patent documents in many countries worldwide (INPADOC)
- European patent register data: Bibliographic, legal and procedural information on published European and EURO-PCT applications (REGISTER)
The EPO's PATSTAT product line: Added value on Applicant names

- **Harmonization of names** to increase usability and connectivity with other data sources

- Applicant and inventor names options in PATSAT:
  - **Original names**
  - **DOCDB** standardized name (twice/year)
  - **HAN** applicant name (OECD) (twice/year)
  - **PATSTAT harmonized name**: Harmonized applicant and inventor names (twice/year)
The EPO's PATSTAT product line: PATSTAT harmonized name

- Procedure (developed by KU Leuven): **Two-step process\(^1\):**
  - methodology layer 1: **automated;** focus on identification of name variations, using data cleaning and processing
  - methodology layer 2: **semi-automated;** increasing completeness, with focus on top 500 applicants

---

# Harmonised Names in PATSTAT

- original name, DOCDB names, externally standardized names

<table>
<thead>
<tr>
<th></th>
<th>original name</th>
<th>DOCDB standardised</th>
<th>EEE_PPAT (Uni. Leuven)</th>
<th>HAN (OECD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TOYOTA MOTOR CORP</td>
<td>TOYOTA MOTOR CORP</td>
<td>TOYOTA MOTOR CORPORATION</td>
<td>TOYOTA MOTOR CORP</td>
</tr>
<tr>
<td>2</td>
<td>Toyota Jidosha Kabushiki Kaisha</td>
<td>NISSAN MOTOR</td>
<td>HONDA MOTOR COMPANY</td>
<td>TOYOTA JIDOSHA KK</td>
</tr>
<tr>
<td>3</td>
<td>NISSAN MOTOR CO LTD</td>
<td>HONDA MOTOR CO LTD</td>
<td>NISSAN MOTOR COMPANY</td>
<td>NISSAN MOTOR CO LTD</td>
</tr>
<tr>
<td>4</td>
<td>HONDA MOTOR CO LTD</td>
<td>TOYOTA MOTOR CO LTD</td>
<td>HYUNDAI MOTOR COMPANY</td>
<td>HONDA MOTOR CO LTD</td>
</tr>
<tr>
<td>5</td>
<td>Hyundai Motor Company</td>
<td>HYUNDAI MOTOR CO LTD</td>
<td>ROBERT BOSCH</td>
<td>NISSAN MOTOR CO LTD</td>
</tr>
<tr>
<td>6</td>
<td>Robert Bosch GmbH</td>
<td>BOSCH GMBH ROBERT</td>
<td>AISIN AW COMPANY</td>
<td>HYUNDAI MOTOR CO</td>
</tr>
<tr>
<td>7</td>
<td>HONDA MOTOR CO., Ltd.</td>
<td>AISIN AW CO</td>
<td>MAZDA MOTOR CORPORATION</td>
<td>BOSCH ROBERT</td>
</tr>
<tr>
<td>8</td>
<td>Nissan Motor Co., Ltd.</td>
<td>MAZDA MOTOR</td>
<td>DENSO CORPORATION</td>
<td>HONDA GIKEN KK</td>
</tr>
<tr>
<td>9</td>
<td>MAZDA MOTOR CORP</td>
<td>FUJI HEAVY IND LTD</td>
<td>FUJI HEAVY IND</td>
<td>DENSO CORP</td>
</tr>
<tr>
<td>10</td>
<td>HONDA GIKEN KOGYO KABUSHIKI KAISHA</td>
<td>DENSO CORP</td>
<td>KUBOTA CORPORATION</td>
<td>HONDA MOTOR CO</td>
</tr>
</tbody>
</table>
Reflections for the future

In a digital world, name standardization is superseded by IDs.

Should we consider applicant IDs in the IP community?

Is ECLI (not centralized) ORCID (central registry) a model? or better follow (sparse) national info?

THERE IS NO LOW HANGING FRUIT
Thank you for your attention!

John Potts, Anne-Marie McConnell, Ignacio Muñoz Ozores, Christian Soltmann
European Patent Office

jpotts@epo.org
amcconnell@epo.org
imunozozores@epo.org
csoltmann@epo.org