HARMONISATION OF IP APPLICANT NAMES: THE OECD EXPERIENCE

Hélène DERNIS & Mariagrazia SQUICCIARINI
OECD, Directorate for Science, Technology and Innovation

March 26, 2021
Entity resolution at the OECD 1/2

**STI Micro-data Lab**
- Gather various sources of data
- Curate, link, sort micro-level data
- To inform policy-oriented questions
Entity resolution at the OECD 2/2

- Linking different datasets using **company names**
- Focus on the **IP data** within the STI Micro-data Lab
- Applicable to different sources of company-level names
Making sense of official IP data 1/2

IP rights (Patents, Trademarks and Design rights) are a rich source of information however:

- **No firm identification codes** in IP documents
- **No unique applicant name** in databases, due to:
  - Typing errors;
  - Use of acronyms or short names (e.g. I.B.M.);
  - Translations;
  - Indication of legal forms (A.G., Gmbh, S.A., Int, Inc., Ltd., Oy…);
  - Punctuation signs …

➔ Enrich the data to inform policy
➔ Entity resolution / harmonisation crucial
IP applicants names need to be linked:

- across different IP sources \((\text{patent} – \text{trademarks} – \text{designs})\)
- to company level data e.g. from ORBIS© database

- Analyse the IP bundle of companies;
- Uncover the characteristics of innovative firms;
- Link to entrepreneurship, productivity, innovation, R&D …
Linking applicants to firm level data 1/3

**Imalinker** *(Idener Multi Algorithm Linker)* designed in 2011 for the OECD

- Matching performed on a *country by country* basis;

**Step 1.** Harmonisation of names using *country specific dictionary* *(building on international effort – e.g. Eurostat/KUL: Magerman et al (2009))*

<table>
<thead>
<tr>
<th>Corporation</th>
<th>Corp</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research &amp; Development</td>
<td>R&amp;D</td>
<td>GB</td>
</tr>
<tr>
<td>Gesellschaft mit beschränkter Haftung</td>
<td>GMBH</td>
<td>DE</td>
</tr>
<tr>
<td>Aktiengesellschaft</td>
<td>AG</td>
<td>DE</td>
</tr>
<tr>
<td>Centre National de la Recherche Scientifique</td>
<td>CNRS</td>
<td>FR</td>
</tr>
<tr>
<td>kabushiki kaisha</td>
<td>KK</td>
<td>JP</td>
</tr>
</tbody>
</table>
Linking applicants to firm level data 2/3

Step 2. Linking harmonised names:

• **String matching** algorithms (token/string metric based):
  - **Token frequency balanced** measure;
  - **Levensthein** distance;
  - **Jaro-Winkler** distance;
  - Other...

• **High precision** (thresholds of high matching scores to minimise the false-positive and false-negative);

Step 3. **Manual controls** to adjust the precision & recall.
Step 4. Disambiguation rules to cope with **one-to-many** matches

1. Control for the **address** of applicants;
2. Exclude links to **branches** or **affiliates**, where several;
3. Identify the **domestic ultimate owner** when available;
4. Exploit business register data to select the “**best match**” when steps 1-3 failed:
   - **Select companies according to the turnover or employment figures**;
   - **Select companies for which financial information is available**;
   - **Other...**
Examples of OECD linked IP data

Degree of precision depends on end usage

R&D investment, IP bundle and publications of the World top 2000 R&D Investors

Consolidation of IP portfolio at firm level

**Caveats** to precision / recall:

- Data **truncation**
  (e.g. *business-register data only contains recent records*)
- **Change** in company names or in group structure
  (*longitudinal analysis*)
- Name **translations**
  (*dealing with non-latin characters*)
- ...

OECD HAN database *(Harmonised Applicant Names)* derived from the name harmonisation procedure and/or on the outcomes of the matching for *patents* (source: EPO PATSTAT database).

- 2 levels of harmonization:
  - **HARM_ID**: Unconsolidated grouping – more conservative applicant names grouped together using the same “cleaned name”
  - **HAN_ID**: Consolidated grouping applicant names grouped together using both “cleaned name” and link to business register data

- A *single harmonised name* is proposed per each HARM_Id or HAN_Id (based on the most frequent name identified within each grouping)

- Most OECD countries + others are covered.
Reduction of name variation (PATSTAT)

Reduction by country

Entity resolution of applicant names is key to:
• Better assess the IP portfolio of applicants;
• Improve the relevance of data;
• Facilitate linkages across data sources.

Further information:
- OECD work on IP Statistics: [oe.cd/ipstats](http://oe.cd/ipstats)
- Free access to OECD HAN database
- IP related statistics published in OECD reports and on OECD.Stat portal
- [sti.microdatalab@oecd.org](mailto:sti.microdatalab@oecd.org)