OECD HAN DATABASE

HARMONISED APPLICANT NAMES - *February 2022 edition*

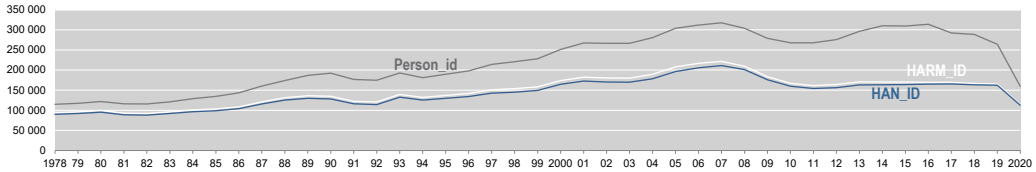
BACKGROUND INFORMATION

The OECD HAN database provides a grouping of patent applicant's names which has been elaborated with business register data. The names of patent applicants were originally extracted from *European Patent Office's (EPO) Worldwide Statistical Patent Database (PATSTAT, Autumn 2021)*. The database also includes the list of patent documents filed to the EPO, the US Patent and Trademarks Office (USPTO) or through the Patent Co-operation Treaty (PCT).

METHODOLOGY

The groupings of patent applicant names (PATSTAT's PERSON table) have been performed as follows:

- **Cleaning and harmonising**: names are corrected from punctuation, accents, abbreviations and legal information, using dictionaries developed on a country basis. A preliminary grouping is generated upon the harmonised name. A new grouping based on the first level of harmonisation is provided as a new **HARM_id**.
- **Consolidating**: cleaned/harmonised names were matched against company names from business register data (as provided in the *ORBIS© database* from Bureau van Dijk Electronic Publishing, March 2020). The matching was performed using series of algorithms (*approximate string matching; weighted token-based comparisons; distance measures*) within the *IMALINKER* system developed for the OECD by IDENER, Seville 2013. Each algorithm computes a matching score per pair of names, assessing therefore for the likelihood of names similarity. The matched pairs of names are selected according to high thresholds of matching scores in order to maximise the precision of the match. Finally, names are further grouped together according to either the matched ORBIS© company name or the cleaned/harmonised names resulting from the algorithms. The harmonisation of names was propagated to new applicant names from the latest edition of PATSTAT's PERSON table.
- **Grouping:** A unique identifier **HAN_id** is automatically generated for each grouping of patent applicants. A common name is then attributed to each HAN_id group according to the first applicant of the grouping (name of the applicant that contributed to the highest number of patent applications).



Due to the large volume of data processed, it was not possible to control each names grouping. Errors may therefore be encountered: any feedback on incorrect harmonisation would be highly appreciated.

DATASET COVERAGE

The OECD HAN database, February 2022, provides groupings of patent applicant's names for most OECD countries and countries in the BRIICS. The list of patents filed to the EPO, the USPTO and through the PCT is made available for each grouping of applicants. Further improvements are expected in future versions, notably on the countries coverage.

RESTRICTIONS, SOURCE & CONTACT

Please note that the OECD HAN database is provided for research and analytical work. When publishing the results of your analysis, make sure it is quoted as: ***"OECD, HAN database, February 2022".***

For further information about OECD patent related work, methodologies and access to patent indicators, please visit our web page at: oe.cd/ipstats.

Comments and questions about this dataset should be sent to STI.Microdata@oecd.org.
For further information on EPO's PATSTAT, please contact patstat@epo.org.

DATABASE STRUCTURE

The OECD HAN database, February 2022, consists of 4 distinct tables presented in flat files (UTF-8 format, columns separated using the pipe "|" character). Applicant's idenfiers from the last editions of PATSTAT are linked to ***4,163,793*** unique HAN_id, and ***4,549,598*** unique HARM_id.
Note that changes in the identifiers may occur from one version to the next.

HAN_PERSON
Correspondance table between HAN_id , HARM_id and Person_id — ***7,674,777 rows***

HAN_id	Unique identifier - grouping based on similar names and links to company level data *Modified at each data release*
HARM_id	Unique identifier - grouping based on similar names only *Modified at each data release*
Person_id	Applicant identifier from PATSTAT, Autumn 2021
Person_name_clean	Harmonised applicant name
Person_ctry_code	Applicant's country
Matched	Indicator of sucessful match to ORBIS© (=1 if matched)

HAN_NAMES
Harmonised names associated to each HAN_id — ***4,163,793 rows***

HAN_id	Unique identifier - grouping based on similar names and links to company level data *May be modified at each data release*
Clean_name	Proposed harmonised name (top applicant name in the HAN grouping)
Person_ctry_code	Applicant's country

HARM_NAMES
Harmonised names associated to each HARM_id — ***4,549,598 rows***

HARM_id	Unique identifier - grouping based on similar names only *May be modified at each data release*
Clean_name	Proposed harmonised name (top applicant name in the HARM grouping)
Person_ctry_code	Applicant's country

HAN_PATENTS
Patents filed by each HAN_id (for EPO, USPTO, PCT only) — ***18,009,199 rows***

HAN_id	Unique identifier - grouping based on similar names and links to company level data *May be modified at each data release*
HARM_id	Unique identifier - grouping based on similar names only *May be modified at each data release*
Appln_id	Surrogate key - patent application identifier in PATSTAT, Autumn 2021
Publn_auth	Publication authority
Patent_number	Patent publication number - normalised format EP*XXXXXXX* (patent published by to the EPO) US*XXXXXXX* (patent granted by USPTO) US**YYYY***XXXXXX* (patent published by USPTO) WO**YYYY***XXXXXX* (publication of patent application filed through the PCT) where **YYYY** represents the filing year and *X* in {0-9}