

NEW PATENTS DATABASES WITH HARMONISED APPLICANTS' NAMES AVAILABLE TO RESEARCHERS

EUROSTAT - OECD

In the framework of the **OECD Task Force on Patent Statistics**,¹ **EUROSTAT** and the **OECD**, together with academics from the Katholieke Universiteit Leuven (Belgium) and from University of Camerino (Italy), engaged in a co-operative work for developing methods and building patents databases with cleaned applicant's names which can be matched with business register databases. The two databases are complementary to each other: the KUL/Eurostat database is based on in-depth cleaning and harmonisation of the name of applicants; the OECD database has used the names of companies as reported in business registers as the reference for harmonisation.

1. KUL/EUROSTAT: Method for harmonising applicants names and PATSTAT harmonized PERSON table

The K.U.Leuven/EUROSTAT method for harmonized patent applicant's names is a comprehensive method to achieve harmonization of patentee names in an automated way. This method was applied on all applicant names in the PERSON table of EPO's Worldwide Patent Statistical Database (PATSTAT), edition April 2009, resulting in a new PERSON table with harmonized applicant names. The developed method is based on the contents of the name and country address of the applicant name, no other information is used in the harmonization process. All names are processed by a step-wise validation process based on rules: character cleaning; punctuation cleaning; legal from indication treatment; common company word removal; spelling variation harmonization; condensing; umlaut harmonization. About 4,000 *search and replace* rules are executed for every step to handle the particular issue.

A first version of the method was applied on EPO and USPTO applicant names in 2006 (EUROSTAT Working Paper - Magerman, T., Van Looy, B. & Song, X. (2006) Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. The method was extensively updated in 2009 based on all PATSTAT (04/09) applicant names (~11.000.000). The number of rules was extended to almost 4000 to improve recall without jeopardizing precision. This current beta version is used to create a harmonized person table with the PATSTAT PERSON_ID, original name and harmonized name for easy linking to the existing PATSTAT PERSON table (edition April 2009). The data is available upon request for research (contact tom.magerman@econ.kuleuven.be). A final version will be published when the October 2009 version of PATSTAT will be released, together with a working paper with the description of the method and results and impact on the PATSTAT applicant names. All publicly available material will be disseminated through EUROSTAT web site.

2. OECD: HAN database

The OECD HAN ("Harmonised Applicants' Names") database provides a dictionary of applicants' names which have been elaborated with business register data, so that it can easily be matched by all users. It results from a three steps data processing: (1) Identify business organisations, non business organisations and individuals among patent holders; (2) Clean the company names; (step 1 and 2 based on KUL algorithm – see above); (3) Consolidate the cleaned names by matching patent data with other databases, e.g. business register.

The beta-version of OECD HAN database proposes a "dictionary" of names where consolidated harmonised names were regrouped according to a unique identifier. OECD HAN table can be used together with PATSTAT (04/09). The data is accessible to researchers that are conducting further analysis at the micro-level, with the aim of benefiting from user feedback on any inconsistencies that would be identified in order to improve it in the next versions. A matching software is going to be made available to researchers so that they can link the patent data to business registers or other business databases. Methodological documentation on HAN database: Thoma et al (2009) – Harmonizing and Combining Large Datasets – An Application to Patent and Finance Data forthcoming STI Working Paper, OECD. Access to the data should be requested to sti.contact@oecd.org (mentioning *OECD, HAN database* in the title of the message).

¹ The OECD Task Force on Patent Statistics gathers representatives from Eurostat, the European Patent Office (EPO), the Japan Patent Office (JPO), the US National Science Foundation (NSF), the US Patent and Trademark Office and the World Intellectual Property Organization (WIPO).