OBSERVATOIRE DES SCIENCES ET DES TECHNIQUES
Au service de tous les acteurs de R&D

Les notes pratiques de l'OST

Matching patent data with financial data[1]

An organization's patent portfolio forms a critical part of its assets and may greatly influence its strategy and market value. Nevertheless, even if recently patent offices have changed their policy on data to make them easier to access, most patent data users are still unable to identify non-ambiguously the patent applicants[2].

One of the main reasons is that patent data are collected to identify the novelty of a invention and to make it public as a new stock of knowledge for further inventors, whatever the applicants. The process of the quality of information is then, first of all, focused on the data regarding scientific and legal information. Thus for instance, every patent has an unique ID number in order to identify perfectly the invention and to be able to establish scientific links between patents. By contrast, no quality check is applied to applicants' names and addresses, and in some cases (US data for instance) apart from the country even address data are unavailable. Thus identifying (called afterwards disambiguating) existing applicants by name or address, in order to build up a unique identifier for each patenting entity, is not a simple task.

Another issue may also be the **timeframe**: patent attributes are usually a snapshot of data at the moment the dataset producer (for instance EPO or USPTO) releases them. If the producer does not receive updates (for example because this is not required from applicants or because the data producer has no need of such data[3]) such attributes are *frozen* at the moment of last update. Patents granted 10 years ago will, in some cases, have been applied for by expired/split/merged/ acquired entities. For instance, it may not be possible to assign patents owned by Compaq to Hewlett Packard, by which it was acquired in 2002, since the patent might still be in the name of Compaq.

Last but not least patent data do not include ***applicant group structures***, so using them alone it is not possible to consolidate patent portfolios by "Global Ultimate Owner" (GUO).

For such reasons a third party data source is needed, containing for example company history and structure.

As a matter of fact there are currently several existing data sources containing financial data, indicators, private equity data and portfolio organized by company, where company ownership structure is also available, as well as their history in terms of mergers and acquisitions, name changes or other events that impact on their structure.

The purpose of this document is to illustrate the algorithm that we are developing in order to match patent and financial data through a general purpose methodology that may also be applied when reconciling other data sources.

Other attempts had previously been made. For instance, Grid Thoma et al. (2010) described a methodology to match Amadeus [EU companies] and Patstat for EPO and USPTO patents using a powerful string comparison methodology. In our project we have used these previous efforts, but we extend the match scope to all application authorities and all companies contained in ORBIS, even if extensive usage has been restricted to three of them (EPO, USPTO and INPI). We have also focused on particularly high-tech companies, selected from the industrial sector using NACE 2.0 aggregation[4] for the purposes of the research project in which this algorithm has been developed. We also enrich our methodology by adding to string comparison tools extensive usage of other data from both datasets in order to remove false positive matches. The method that we are developing is characterized by three steps: harmonization, match and filtering. These will be illustrated in the following paragraphs. Before going further, we describe the data sources used.

---

[1] This note is a final comment on the results of a project named 'Valeur Brevet' carried out with Emilie-Pauline Gallié, Lorenzo Cassi, Anne Plunket, Michele Pezzoni and Valérie Mérindol.

[2] Defined as a '*partnership, corporation, or other organization having the capacity to negotiate contracts, assume financial obligations, and pay off debts*'. (http://www.w3.org/2009/03/xbrl/naming.html)

[3] Patent offices like EPO or USPTO usually cease data collection when they grant the patent. This means for instance that after the patent is granted changes in applicant name or patent ownership are not necessarily reported in the database.

[4] Defined in http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/High-tech_statistics

Data sources used

The primary source of information about patents is **PATSTAT**[5], produced by the European Patent Office (EPO), OECD and Eurostat. PATSTAT stands for EPO Worldwide Patent Statistical Database. It has been created for use by government/intergovernmental organizations and academic institutions. It contains a snapshot of the EPO master documentation database (DOCDB) which covers more than 90 national and international patent offices.

Data include *anagraphic* data (such as application date and technology field), forward and backward citations, and family links. This database is designed to be used for statistical research and requires the data to be loaded into the customer's own database (the data are essentially the same as DocDB, but fields have been added to make it easier to retrieve statistical data such as additional address information extracted from US and EP registers).

This database is currently the reference for the calculation of patent indicators, both in the academic world and for public policymakers. It is, among others, used by the OECD for the production of technology indicators.

The source of information for financial data and group identification, including subsidiaries is **ORBIS**[6], from Bureau Van Dijk. ORBIS compiles financial information on nearly 100 million public and private companies worldwide. Among the wide range of information proposed, particularly that related to company subsidiaries, shareholders and industrial sectors have been used as complementary information, along with company names and country.

Dataset harmonization

The quality of the data sources is extremely varied: while ORBIS data have been to a certain extent normalized, PATSTAT applicant names are very raw. Company names may appear with different spellings: the typical example is IBM that may appear also as International Business Machines, but also as I.B.M., Int Busn Mach. This makes it difficult to automatically combine patents applied for by the same entity under a single label.

In order to overcome this difficulty, we harmonized the datasets in two steps:

- We decided to use the ECOOM-EUROSTAT-EPO PATSTAT Person Augmented Table (EEE-PPAT)[7] that uses a wide-ranging method to obtain harmonized patentee names automatically (described in detail in Magerman et al. (2009)). The EEE-PPAT table adds an important piece of information: assignee sector allocation. This identifies whether patentees are private business enterprises, universities/higher education institutions, government agencies or individuals. In this way we have excluded from our exercise applicants flagged as individuals. This reduced the number of applicants in the dataset.
- We processed the data sources comprehensively in order to remove nonstandard ASCII characters, double spaces, and other common misspellings/typos and remove the legal designation (LLC, INC, CORP…). This information was stored in a different field for future in case it may be useful for further disambiguation. The same processing algorithm was applied to both datasets in order to ensure the results can be compared.

After dataset harmonization, our population contained 336 277 distinct entities from Orbis and 3 962 683 entities from Patstat.

Matching the data

Due to the high percentage of standard names, and the fact that in case of company names there may be only small differences between distinct companies (since most of them are acronyms[8]), we decided not to use any edit distance criteria (like Levensthein or N-gram functions) but to rely on three criteria[9] to rank types of match:

1) **Perfect match**: where names, apart from legal designation, are exactly the same;
2) **Alphanumeric match**: where the names are the same, when only [A-Z] and [0-9] are taken into consideration (e.g.: I.B.M. = IBM = I B M);

[5] More info at http://www.epo.org/searching/subscription/raw/product-14-24.html

[6] See http://www.bvdinfo.com/Products/Company-Information/International/ORBIS.aspx

[7] More at http://www.ecoom.be/nl/EEE-PPAT

[8] multi-screen cinemas vs AMD – Advanced Micro Devices

[9] The term edit distance refers to the number of operations needed to transform a string into another. For instance to transform CAR into RARE involves replacing C by R (CAR à RAR) and the insertion of a trailing E (RAR à RARE) giving an overall edit distance of 2.

3) **Jaro Winkler**[10] **Token similarity**[11]: names are broken into tokens and the similarity score is computed by the number of tokens in common, weighted in inverse proportion to frequency. Only results above a threshold value will be considered valid matches.

In case of multiple types of match, only the best result will be considered as a match, where a perfect match is better than an alphanumeric match, which is in turn better than a token similarity. We introduced this condition to avoid the match of non homogenous entities (see the example below where we match a main company and subsidiaries; in most cases, subsidiary names are equal to main company name plus something, and so are more likely to result in a token similarity match rather than a perfect or alphanumeric based).

Example

Names in PATSTAT
P1: ASEA BROWN BOVERI
P2: ASEA-BROWN-BOVERI POWER PRODUCTS

Names in ORBIS
O1: ASEA BROWN BOVERI
O2: ASEA BROWN BOVERI POWER PRODUCTS

MATCH PHASE
P1: matches O1 (perfect match) and O2 (token similarity since power and products are very common tokens) à only best match remains so P1 matches O1
P2: matches O2 as alphanumeric match only.

Aside from name, also country code is taken into account as a main criterion for matching, because the same applicant name in different countries may refer to distinct entities (for instance Ministry of Health). Since in PATSTAT, apart from EPO, country code coverage is sometimes very poor, this data is also enriched using priorities and family data (where in a patent family a homonym with a non-blank country code exists, such data is retained for the examined patent).

Note that, since the timeframe of the two datasets is different (ORBIS contains the latest data available, PATSTAT uses names at the moment of the application or granting of the patent), previous ORBIS names and 'also known as' (AKA) names have also been included in the match, when available in the ORBIS dataset.

At the end of the matching process, the algorithm returned the following figures by type of match:

Table1: results of the matching

Match type	No. of matches
1	43 058
2	19 620
3	1 388 777

Figures above are in the expected range, especially for type 1 and 2 matches, because we would expect only some of the 336 277 entities in ORBIS to be also active in patenting activity; it should be noted that the number of "type 3" matches has been left as loose as possible since names in the two datasets are normalized using different standards, leaving to the next step of the algorithm (filtering phase) the burden of deleting false matches, which we expect to be the majority of type 3 matches.

Filtering the match

The Filtering phase of the algorithm aims to delete false matches and to fine-tune parameters (for instance, threshold value for token similarity). Whenever trying to match two distinct data sources, an inevitable tradeoff arises between flagging positives matches as false negatives and including false positives as true matches. In the first case an existing link between a patenting and a financial entity would not be established, in the second case a non-existing link would be established.

In computer science, this is measured by two variables: precision and recall defined as:
Recall rate = true positives / (true positives + false negatives)
Precision rate = true positives / (true positives + false positives)

Ideally both values should be as near as possible to one, but in reality these rates are used to balance filtering criteria in order to allow the most satisfactory solution among those available.

For this purpose, a manual check was carried out on a control sample made up of 1% of the total ORBIS population involved in the match (about 3500 company names randomly selected),

[10] Jaro Winkler metric is well described in Wikipedia: http://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance

[11] Note that criteria 3 includes matches already retrieved with type 1 (perfect match) but this distinction was made in order to optimize computing time by excluding type 1 and 2 matches from the type 3 search algorithm.

looking for them on ESPACENET[12] (online patent database provided by EPO) in order to obtain an accurate specimen to be used as a benchmark for the overall algorithm. The results of this crosscheck were used to balance the algorithm.

Other criteria were occasionally used to highlight possible false positives. Such criteria included:

a) same technology fields (introducing a cross check via a IPC- NACE concordance in order to highlight matches where patents are in a totally different field from company NACE),

b) timeframe (crosscheck application year and company start-up year on ORBIS),

c) other relevant ORBIS data (e.g. R&D expenses > 0).

Match Results

After filtering, the algorithm described produced 133 554 matches, including 94 661 patenting entities and 66 234 financial entities.

We created a control sample of ORBIS companies classified under NACE 2651 (Manufacture of instruments and appliances for measuring, testing and navigation) to validate and analyze the match. We selected this industrial sector for its high patenting propensity, and manual checks on the sample produced a precision rate of 78% and a recall rate of 91% across all application authorities in Patstat. Table 2 highlights the major contributions to precision and recall, and an analysis of errors by cause.

The most common errors were due to a missing country code in Patstat data, and differences in language between Patstat and Orbis (e.g. 'EvU - Innovative Umwelttechnik' in Orbis as against 'E.V.U. ENTWICKLUNG VON UMWELTTECHNIK' in Patstat).

Our methodology introduces two useful innovations: first, in terms of precision, the addition of 'Also Known As' names from Orbis to the name pool (this was expected because the more alternative names we have, the more matches we generate); second, in terms of recall, checking application dates against foundation date (thus excluding matches where patents had been issued long before the company foundation date).

Table 2: contributions to results

	Error contribution (% of total errors)	Error contribution (% of all matches)	Contribution to precision	Contribution to recall
Different language	45.50%	9.10%		
Missing Patstat country code	40.50%	8.10%		
Token sensitivity too low	11.00%	2.20%		
Other	3.00%	0.60%		
Name/country comparison (type 1,2)			42.75%	41.30%
Jaro Winkler Token similarity (type 3)			52.25%	50.50%
Checking application date against foundation date				8.20%
Adding 'also known as' names			5.00%	

When we restrict the analysis to results from EPO, USPTO and INPI, the results are far better because we eliminate many mistakes due to poor data quality (especially those labeled as 'different language' and 'missing country') thus obtaining a **precision of 91% and a recall of 95%** that are within the expected target.

[12] http://worldwide.espacenet.com

Figures of matched entities show that it is possible to further disambiguate the data by collapsing distinct patenting entities linked to the same financial entity and vice-versa. Another improvement that may be applied in the future is the addition of information from datasets on deals/mergers (one example may be Zephyr[13] also from Bureau Van Dijk) in order to use ownership information and add names of expired/acquired companies.

Last but not least, the algorithm produced can be broadened without modification by adding other industrial sectors to the ORBIS pool of companies or, with minor modifications, be used for matching other data sources. ■

References

Grid Thoma, Salvatore Torrisi, Alfonso Gambardella, Dominique Guellec, Bronwyn H. Hall, Dietmar Harhoff Harmonizing and combining large datasets – An application to firm-level patent and accounting data; 2010; NBER working paper series 15851; http://www.nber.org/papers/w15851

Magerman T, Grouwels J., Song X. & Van Looy B. (2009). Data Production Methods for Harmonized Patent Indicators: Patentee Name Harmonization. EUROSTAT Working Paper and Studies, Luxembourg

[13] http://www.bvdinfo.com/Products/Economic-and-M-A/M-A/Zephyr

Cette note pratique a été rédigée par Gianluca Tarasconi

Ce document a bénéficié de l'assistance éditoriale d'Emilie-Pauline Gallié et de Marie-Laure Taillibert

Correspondance : emilie-pauline.gallié@obs-ost.fr