Huberty, Mark; Serwaah, Amma; Zachmann, Georg

Working Paper

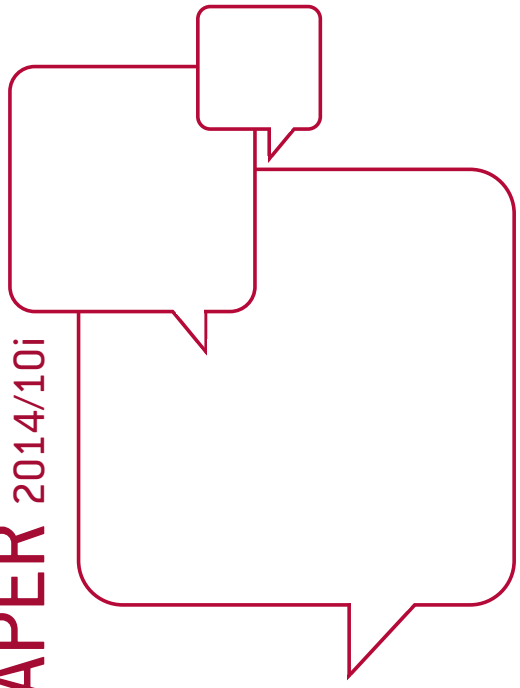
A flexible, scaleable approach to the international patent 'name game'

Bruegel Working Paper, No. 2014/10i

Provided in Cooperation with:
Bruegel, Brussels

Suggested Citation: Huberty, Mark; Serwaah, Amma; Zachmann, Georg (2014) : A flexible, scaleable approach to the international patent 'name game', Bruegel Working Paper, No. 2014/10i, Bruegel, Brussels

This Version is available at:
https://hdl.handle.net/10419/126714



BRUEGEL WORKING PAPER 2014/10i

A FLEXIBLE, SCALEABLE APPROACH TO THE INTERNATIONAL PATENT 'NAME GAME'

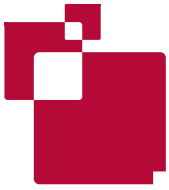
MARK HUBERTY*, AMMA SERWAAH** AND GEORG ZACHMANN†

Highlights

- This paper reports a new approach to disambiguation of large patent databases. Available international patent databases do not identify unique innovators. Record disambiguation poses a significant barrier to subsequent research. Present methods for overcoming this barrier couple *ad-hoc* rules for name harmonisation with labour-intensive manual checking. We present instead a computational approach that requires minimal and easily automated data cleaning, learns appropriate record-matching criteria from minimal human coding, and dynamically addresses both computational and data-quality issues that have impeded progress. We show that these methods yield accurate results at rates comparable to outcomes from more resource-intensive hand coding.

* University of California, Berkeley; markhuberty@berkeley.edu
** Bruegel
† Bruegel; georg.zachmann@bruegel.org

This work was supported by the SIMPATIC project (EU Seventh European Union Framework Programme, grant agreement no. 290597). Mark Huberty received additional support from the United States Environmental Protection Agency STAR fellowship for work relating to the political economy of green innovation. We thank in particular Lee Fleming for use of his disambiguation engine and helpful consultation; Forest Gregg and Derek Elder, who wrote the dedupe library for Python; and KU Leuven, for providing the benchmark data based on their disambiguation of PATSTAT. All errors remain our own.



SEPTEMBER 2014

1 Introduction

Interest in innovation micro-data has risen along with the availability of detailed patent databases. However, data quality poses significant barriers to their use. Patent databases do not reliably identify all instances of the same inventor as a unique entity. Hence the "name game"–correctly identifying which patents belong to the same inventor–has received significant interest. Enabling accurate name disambiguation that scales to the millions of legacy patents and the thousands of new patents filed each year poses both conceptual and computational challenges.

We illustrate an approach to the name game that disambiguates innovator names rapidly, at high rates of accuracy, with relatively minimal human intervention. For all but the most prolific innovators, disambiguation of hundreds of thousands of innovators can be done in only a few hours on consumer-grade hardware. The approach is implemented in an open-source library in a high-level language, permitting relatively straightforward experimentation and customization. We suggest that this approach has significant promise for resolving some of the major challenges in the name game going forward.

2 Prior work on PATSTAT disambiguation

The PATSTAT database, supplied by the European Patent Office, consolidates patents from over 80 individual patent offices worldwide.[1] As of the October 2011 release, it contained upwards of 41 million inventors associated with 73 million patent applications. Regular releases and expanding coverage have made it a valuable resource for research on cross-national innovation.

[1] See http://www.epo.org/searching/subscription/raw/product-14-24.html for more information on PATSTAT.

PATSTAT does not provide unique identification of individuals in its database. Instead, one individual may be identified under more than one identity, depending on how their patent office manages inventor registration. Those identities may share the same name, but under different numerical identifiers. Or, their names may vary in spelling (either alternate spellings or misspellings), the presence of legal identifiers for firms, the order of name components, and other aspects. Hence researchers must play the "name game", wherein they attempt to identify all instances of an individual and associate them with a single unique identifier.

Solutions to the "name game" tend to fall into two categories. The first, widely employed for PATSTAT, uses ad-hoc find-and-replace rules to harmonize names as much as possible. Approximate string matching then attempts to identify all valid similar names from the resulting pool of "clean" names and associate them to a unique identifier. This approach was pioneered by Raffo and Lhuillery (2009), and is presently used to generate the OECD Harmonized Applicant Name (HAN) database. Additional manual disambiguation may further improve the accuracy of the matched data. The KU Leuven variant on HAN undertakes manual disambiguation on prolific inventors. That variant reports precision and recall rates exceeding 99% for the largest innovators, when measured against a small hand-curated set of patents (Callaert et al., 2011).

A second approach applies statistical and machine learning techniques to cluster equivalent names together. Instead of using only name data, as in the ad-hoc string cleaning approach, machine learning methods disambiguate based on a broader profile of the inventor. Lai et al. (2011) treat the record of interest as the inventor-patent instance. Inventors are thus ultimately identified by both personal information (name and address) and patent information (coauthors and technology categories). These multidimensional profiles are then amenable to a semi-supervised approach to statistical matching, drawing on

the Bayesian approach implemented by Torvik et al. (2005), that clusters together patents associated with the same inventor.

Both approaches have their drawbacks. The ad-hoc approach focusing on name alone ignores valuable information about innovators that may improve results. For instance, the "John Smith" patenting in agricultural technologies is likely not the "John Smith" patenting in thin-film semiconductors. Incorporating such additional information can help disambiguate records more accurately. Furthermore, supplementing this manual approach with human intervention requires a significant and ongoing commitment of resources.

Machine learning resolves these issues to some extent, but as implemented depends on reasonably well-formatted data to function appropriately. The United States Patent and Trademark Office data used by Lai et al. are very well-curated. Names are reliably separated into first, middle, and last components; and address information is complete in most cases and is reliably separated into street, city, US state, and country. In contrast, the name data in PATSTAT are not reliably formatted, and address data are both sparse and inconsistent. Figure 5 shows that addresses are sparse for most countries in the European Union. Of those addresses, quality can vary from a complete street address with city and postcode, to only city. In other cases, the address is missing, but can be found in the inventor name field on closer inspection. This variability results, as figure 6 shows, in significant variation in machine-readable and parse-able addresses.

This paper illustrates an new approach that attempts to correct for the shortcomings of both methods as applied to PATSTAT. We implement a machine learning approach to inventor disambiguation. Our approach uses more flexible rules for grouping and aggregating data than the approach taken by Lai et al. (2011), permitting greater accuracy on poorly-structured patent data. We further illustrate how this approach can be structured

to accurately aggregate both high-frequency innovators with significant but unimportant name variance; and low-frequency innovators where name variance is highly important to distinguishing unique inventors. We show that these methods approach levels of accuracy obtained by hand-matching, at a fraction of the time required.

3 Data

The PATSTAT data are characterized by extreme cross-national variability in data quality and completeness. Figure 5 illustrates how variable record completeness can be for EU-27 countries. In general, PATSTAT patent profiles reliably include:

- Person or company name
- Patent technical categories (IPC codes)
- Patent / inventor relationships

Inventor names, while complete, have a variety of problems. First, no reliable resolution of name into its components (first/middle/last) is provided. Second, name data often contain unrelated information, such as addresses. Third, some records appear to put the inventor name in the Coauthor field, and vice-versa. Fourth, some records appear to put multiple co-patenters into a single Name. Fifth, non-English names are not consistently transcribed into English phonetics. This is a minor problem for some countries, and a major difficulty for others. For instance, Germany may transcribe an umlaut, as in “Schön”, to either ”Schon” or “Schoen”. Transformations of Asian names into Latin characters generates even more spelling and phonetic variance.

Other data are even less reliable:

- Address data are readily available for US innovators but sparse besides

- Some address data are contained in the name field; but this is not reliably formatted
- Corporate legal identifiers are not consistently formatted in name fields, and are not provided separately

Address data are particularly problematic here. Given two identical names, the geographic location of two inventors may prove critical to determining whether they are the same person. In addition to address sparsity, PATSTAT addresses are not reliably separated from names.The form taken by address data varies from a complete address with street, city, and postcode; to only a city name. This makes address comparison difficult; and makes parsing addresses out of names a complicated exercise.

4 Disambiguation approach

We propose a machine learning-based approach that attempts to address data quality and completeness issues without substantial human intervention. We demonstrate that such an approach can make rapid progress on disambiguation. We present results for country-level innovator data for seven countries in the EU-27: the Netherlands, Belgium, Denmark, Finland, France, Spain, and Italy. For all but the largest countries (France, Germany, and the United Kingdom), disambiguation took less than 3 hours on consumer-grade PC hardware.[2] Precision rates compared with hand-matched Leuven data reliably exceeded 90%, and exceeded 95% for 24 of 25 countries. Recall rates exceeded 95% for 21 of 25 countries. Furthermore, in some cases (as the tables in Appendix C show), our results are arguably superior to Leuven results when examining individual name matches.

[2] All computation used an AMD Phenom II X4 quad-core chip running at 3.2MHz, with 16gb RAM, running Ubuntu 10.04 LTS.

4.1 Overview

Our approach adopts a four-step approach to name disambiguation. All data were drawn from the October 2011 version of PATSTAT:

1. Generate person-patent records with the following data for each PATSTAT `person_id`[3]:
   (a) Inventor name (person or company)
   (b) Address
   (c) Coauthors
   (d) IPC classes
2. Perform basic string cleaning, including case standardization, diacritic removal, and excess whitespace removal
3. Geocode all addresses[4], returning their latitude-longitude pairs. If address data were blank, the name field was checked for address information. If the address was found in the name field, the name and address were split, the record name updated with the address-free name, and the address geocoded.
4. Aggregate all records corresponding to a unique PATSTAT `person_id` into a single person record, consisting of:
   (a) Most common name variant
   (b) Most common non-null latitude/longitude pair
   (c) All unique coauthors, up to a limit of 100

[3]Compare these data to the much more elaborate data available to Lai et al. (2011): first name, middle name, last name, street, city, postcode, country, coauthor, patent assignee, and technical class.

[4]Geocoding was performed using a fuzzy search matching algorithm to match addresses against cities in the Maxmind world cities database (`http://www.maxmind.com/en/worldcities`). See `https://github.com/markhuberty/fuzzygeo` for code and documentation.

(d) All unique 4-digit patent codes, up to a limit of 100

(e) Count of all patents attributed to that `person_id`

Note that if more than 100 coauthors or patent codes were found, 100 were randomly selected for inclusion.

5. Disambiguate the resulting person file using methods for learnable record linkage described by Bilenko (2006) and implemented in the `dedupe` library for Python.[5]

4.2 Disambiguation with learnable blocking and comparison

The Bilenko (2006) approach implemented in the `dedupe` library is particularly attractive for poorly formatted data because it learns *both* the best comparison metric *and* the best blocking rules. All disambiguation methods face an intractable computational problem: the number of possible pairwise comparisons between records scales as the square of the number of records. Comparing all records to each other thus quickly becomes impractical. Resolving this problem requires some form of blocking rule that compares only likely duplicates. Blocking rules are usually rigid–for instance, comparing only those records whose first name shares the same first letter; or grouping together individuals who live in the same city. But the PATSTAT data has both inconsistently-formatted data and missing data, and consequently lacks the standardization that allows rigid blocking rules to perform well.

The `dedupe` library instead learns the best blocking approach from user-labeled data. During disambiguation, `dedupe` presents the user with a stream of potential matched pairs. Pairs are selected to focus on those pairs for which `dedupe` is most uncertain about the match. The user labels these pairs as "match", "nonmatch", or "ambiguous".

[5] See `https://github.com/open-city/dedupe` for more detail on `dedupe`.

The blocker then selects blocking approaches from a stable of possible blocking "predicates" based on this user-labeled data. Predicates go beyond simple heuristics to include more variable options like "any consecutive three letters" or "same latitude/longitude grid cell". This form of blocking can successfully block strings containing the same word entities even if those entities are in different orders. Blocks are constructed from these predicates to maximize the probability that records labeled as matches are placed in the same block together. Using that same labeled data, `dedupe` then learns a set of optimum comparison thresholds for accepting or rejecting matches. By learning both the blocking strategy and the criteria for identifying duplicates, `dedupe` helps account for the two problems created by the variability of the PATSTAT data: the lack of standardized formatting (and hence difficulty of blocking given a pre-specified rule set) and the variation in match criteria owing to cross-national differences in data quality, name homogeneity, and other factors.

Selection of valid distance metrics for computing the similarity of two records requires some care to model name variance for different inventor types. Consider two different types of inventors:

1. Large companies, who patent frequently, under names that display significant–but otherwise unimportant–name variation. E.g., we identify at least four variants on the Dutch firm Philips Electronics: Philips Electronics, Konink Philips Electronics, Philips Gloeilampenfabrik, and Koninklijke Philips Electronics. Spelling and transcription errors create further, unimportant name variance.

2. Individuals patent infrequently, under names common to people from their country of origin. E.g., many distinct American inventors may have the name "James Smith". Relatively minor variance (e.g., between "James Smith" and "Jane Smith") is very important for distinguishing unique inventors.

A valid matching strategy should explicitly model the interaction between the frequency of patenting and other similarity measures. Comparisons between records with similar patent counts should require very close name matches in addition to similarity among non-name features (geography, technology class, and coauthors); comparisons between records with widely divergent patent counts should do the opposite, down-weighting the name comparison and putting greater weight on the non-name features. This weighting scheme will permit large organizations with significant name variation to be lumped together, without also over-aggregating individuals. We implement this with an interaction term between the string distance between names, and the inverse of the differences in patent counts. The complete record distance specification is described in Table 2.

5 Results

We present two results. First, we show that the disambiguation approach described above generates highly accurate results when compared with the Leuven dataset. We note that the Leuven dataset is not, itself, a master record of hand-disambiguated patents. Instead, it too is generated using a set of tuned algorithms, combined with some manual disambiguation. The Leuven dataset provides two levels of disambiguation. All inventors are first consolidated using an extensive process of data cleaning, combined with a form of fuzzy name matching. Of the consolidated inventors, a subset of high-volume inventors are then checked by hand and consolidated further. We will refer to these two degrees of disambiguation as, "Level 1" and "Level 2", respectively. We make use of the Level 2 data in particular as the Leuven results, presented in Callaert et al. (2011), suggest very high rates of accuracy. But the Leuven dataset does not represent a perfect master record, an issue whose implications we discuss further in section 5.1.

Table 1 illustrates how precision and recall were calculated in reference to the Leuven dataset. For precision each `dedupe` unique ID may aggregate PATSTAT IDs associated to more than one Leuven ID. Using the terminology preferred by Lai et al. (2011), we "clump" Leuven IDs together. This corresponds to the formal definition of `precision`: true retrieved matches as a share of all retrieved matches. Conversely, we may also "split" Leuven IDs: the PATSTAT IDs assigned to a single Leuven ID may be assigned to more than one `dedupe` ID. This corresponds to the formal definition of `recall`: true retrieved matches as a share of all retrieved matches.[6]

[6]This method for computing precision and recall reflects a *group-level* view of the matching problem. In that view, the critical metric is how many IDs were assigned to a common group, compared with a ground truth estimate for how should have been assigned. This is not the only metric. Instead, we might have considered a *pairwise* approach, where we wish to know how many IDs were correctly connected to each other, compared with a ground truth. The group-level approach measures group membership, while the pairwise approach measures the connections between IDs represented by group membership. The functional difference is as follows: for a group of size N, to which we assign M IDs, the group-level approach would compute recall as $\frac{M_{correct}}{N}$, whereas the pairwise approach would compute it as $\binom{M_{correct}}{2}/\binom{N}{2}$.

The pairwise approach permits a more nuanced view of success in the following case: consider a group that in reality contained 10 members. We assign 5 members to one group, and the remaining 5 to another. In this scenario, $recall_{group} = \frac{5}{10}$, while $recall_{pair} = \frac{20}{45} = \frac{4}{9}$. By splitting the group in this instance, we have missed many pairwise connections, even though 50% of all IDs were correctly aggregated together.

We choose the group-level approach because it most closely reflects actual user concerns: given a specific dedupe ID, how well does that ID represent the unique inventor composed of all the non-unique PATSTAT records it contains? Group-level membership, rather than pairwise connections, determine the usefulness of our disambiguation for the end user even if it sometimes overstates the rate of correct pairwise association.

PATSTAT ID	Patent Count	`dedupe` ID	Leuven ID	ID Precision	Patent precision	ID Recall	Patent recall
1	5	1	1				
2	1	1	1	$P_{id=1} = \frac{2}{3}$	$P_{id=1,patent} = \frac{6}{9}$	$R_{id=1} = \frac{2}{2}$	$R_{id=1,patent} = \frac{6}{6}$
3	3	1	2				
4	7	2	3	1	1	1	1

Table 1: Stylized example for calculating precision and recall values for `dedupe` in reference to the Leuven dataset.

Field	Distance metric
Name	Learnable affine gap
Lat / Long	Haversine great circle
IPC Code	TfIdf-weighted Cosine set similarity
Coauthor	TfIdf-weighted Cosine set similarity
Patent count	$\frac{1}{\lvert c_a - c_b \rvert + 1}$ for patent counts c in records a, b
Patent $\times$ Name	$Name \times Patent\ count$

Table 2: Distance metrics by field

5.1 Disambiguation of the EU-27

Testing of this single-shot disambiguation with Belgian data suggested that tuning the algorithm to weight recall at 1.5 times precision generated highly accurate results. Application of this setting to countries other than Belgium showed again generated highly accurate results in all but a few cases. As table 3 shows, precision and recall both exceeded 0.95, as compared with the Level 2 data, for cases other than France, Germany, Poland, and Hungary. Moreover, these results were superior to those obtained with the inventor-patent instance approach taken by Lai et al. (2011), while requiring a fraction of the compute time. For all but the largest countries, complete disambiguation took 2-3 hours or less, with at most 20 minutes of human input.

We emphasize that the precision and recall performance mixes two different kinds of potential error: first, we may under-perform Leuven by either failing to find instances of an individual inventor that the Leuven approach did; or by grouping unrelated inventors together. Alternatively, however, we may also do the opposite: finding instances of an individual that the Leuven approach missed, or correctly splitting records that Leuven treated as the same individual. We find both of these in evidence on close examination of the outcomes.

For instance, the table in section C.4, we see that the Netherlands operations of **Schlumberger** are split between its Technology and Holdings groups. All patents by Schlumberger in the Netherlands are thus divided among these subsidiaries. The `dedupe` approach correctly aggregates these patents together; while the Leuven approach keeps them separate. Similarly, in section C.2, Leuven identifies the Lithuanian inventor **Juozas Grazulevivius** as four separate individuals, whereas the `dedupe` approach identifies them as the same. Obvious misses by `dedupe` include splitting variants on the Spanish inventor `Antonio Martinez Martinez` into separate individuals (as shown in section C.1),

and over-aggregating the different faculties of **Ceska Vysoko Ucena Technical v. Praze**, as shown in section C.2. Hence, the performance of the `dedupe` algorithm in reference to the Leuven dataset mixes improvements to the Leuven result, mistakes, and differences in levels of aggregation for companies and their subsidiaries.

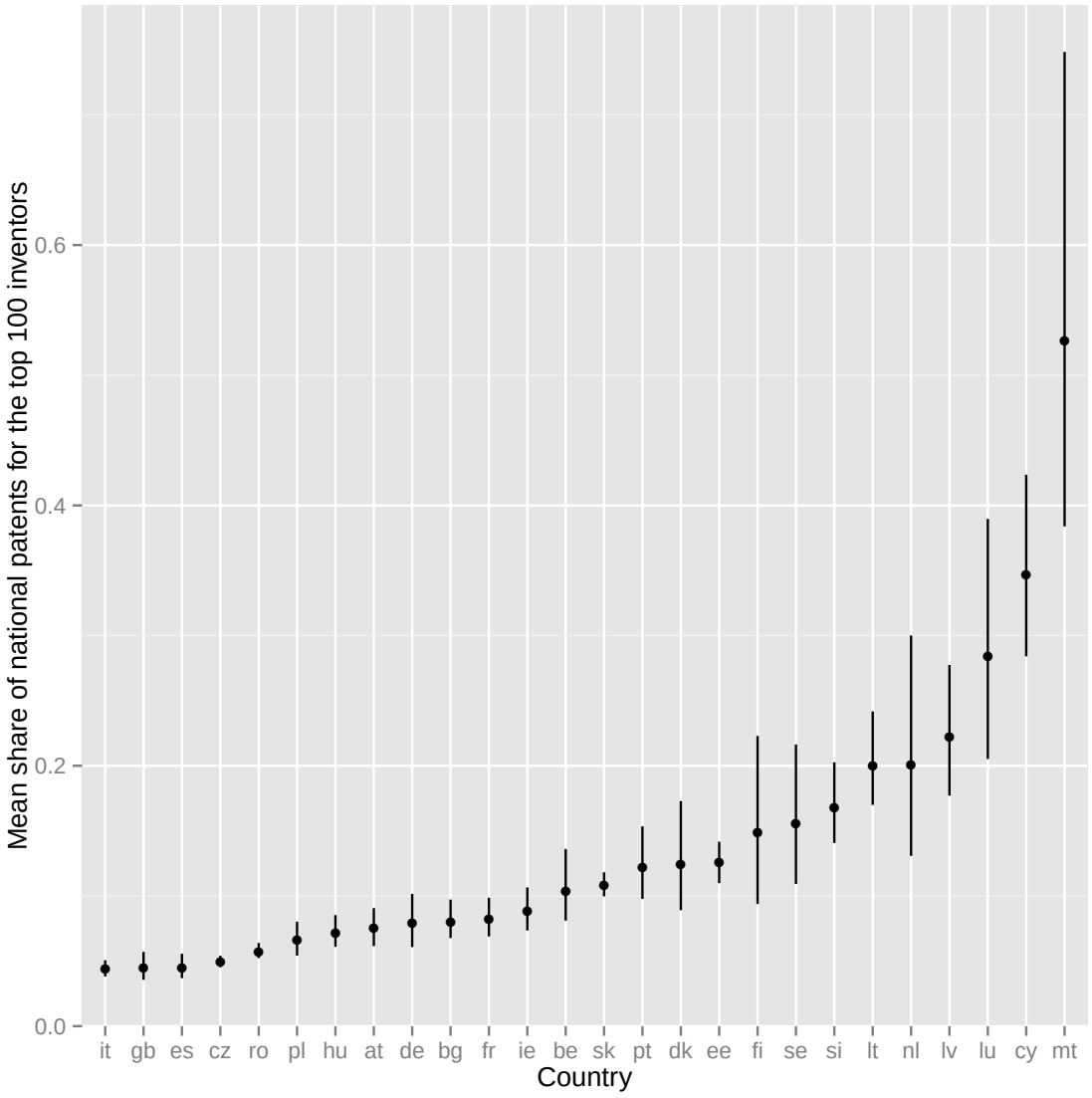


Figure 1: Mean patent share per individual for the top 100 innovators in the raw PATSTAT data.

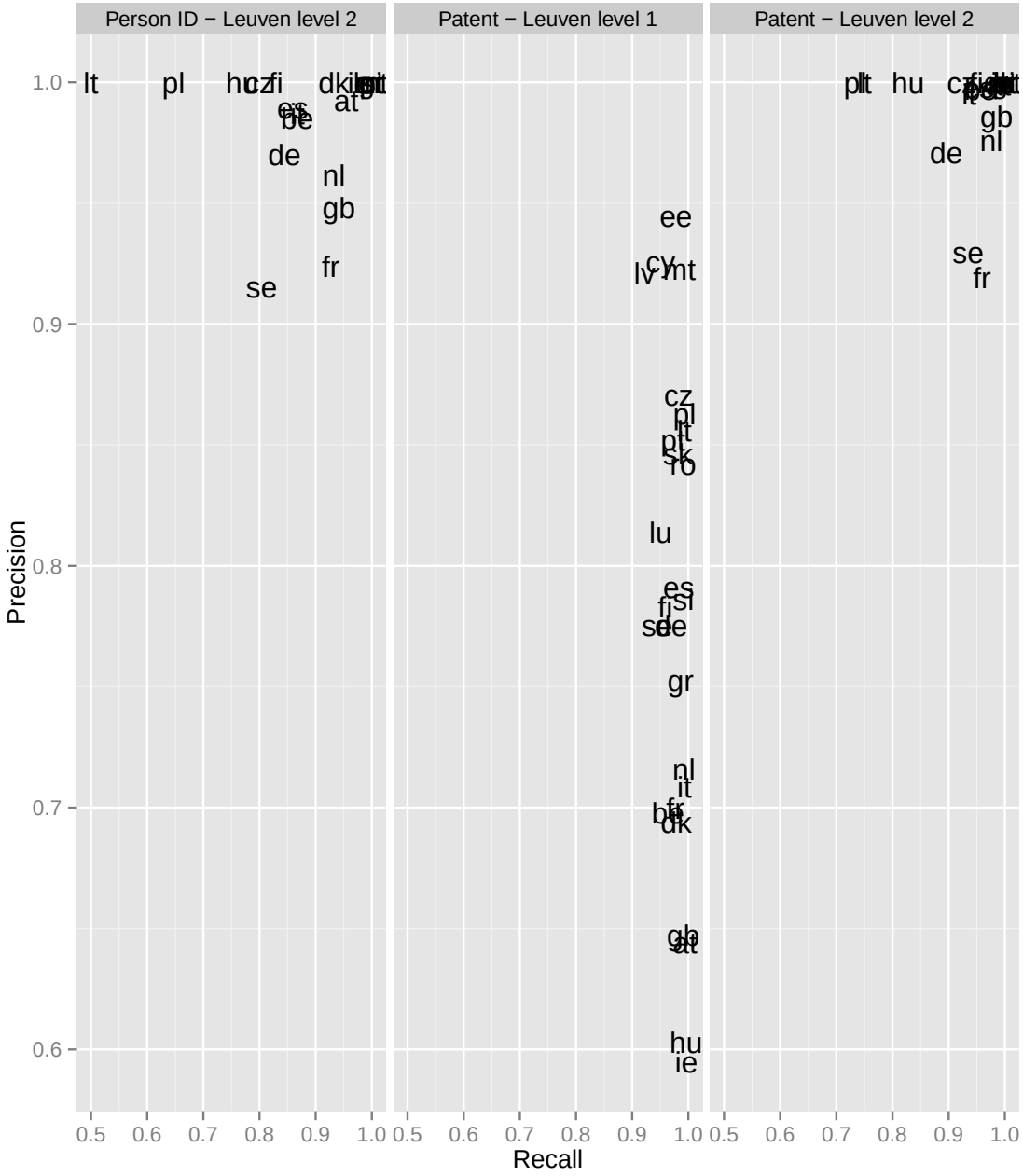


Figure 2: Precision and recall results for seven EU-27 countries across two stages of disambiguation. Id-level values computed in reference to Leuven level 2 IDs. Patent values are relative to the Leuven L1 and L2 IDs. Lines indicate changes in precision and recall from stage 1 to stage 2.

	Country	ID precision	ID recall	L1 patent precision	L1 patent recall	L2 patent precision	L2 patent recall	Precision-Recall weights
1	AT	0.99	0.95	0.64	0.99	1.00	0.99	3.00
2	BE	0.99	0.87	0.70	0.96	1.00	0.96	1.25
3	CY			0.93	0.95			1.50
4	CZ	1.00	0.80	0.87	0.98	1.00	0.92	1.50
5	DE	0.97	0.84	0.78	0.97	0.97	0.90	1.50
6	DK	1.00	0.93	0.69	0.98	1.00	0.99	1.50
7	EE			0.94	0.98			1.50
8	ES	0.99	0.86	0.79	0.98	1.00	0.95	1.50
9	FI	1.00	0.83	0.78	0.96	1.00	0.95	1.50
10	FR	0.92	0.93	0.70	0.98	0.92	0.96	1.50
11	GB	0.95	0.94	0.65	0.99	0.99	0.99	2.00
12	GR	1.00	1.00	0.75	0.99	1.00	1.00	1.50
13	HU	1.00	0.77	0.60	1.00	1.00	0.83	5.00
14	IE	1.00	0.98	0.59	1.00	1.00	1.00	3.00
15	IT	0.99	0.87	0.71	0.99	1.00	0.94	4.50
16	LT	1.00	0.50	0.86	0.99	1.00	0.75	3.00
17	LU	1.00	0.99	0.81	0.95	1.00	1.00	1.00
18	LV			0.92	0.92			1.50
19	MT	1.00	1.00	0.92	0.98	1.00	1.00	2.00
20	NL	0.96	0.93	0.72	0.99	0.98	0.98	3.50
21	PL	1.00	0.65	0.86	0.99	1.00	0.73	6.00
22	PT	1.00	1.00	0.85	0.97	1.00	1.00	1.50
23	RO			0.84	0.99			1.50
24	SE	0.92	0.80	0.78	0.94	0.93	0.93	1.00
25	SI	1.00	1.00	0.79	0.99	1.00	1.00	1.50
26	SK			0.85	0.98			1.50

Table 3: Precision and recall results by country for the `dedupe` output. ID values measure the person-level performance, relative to hand-matched Leuven Level 2 results. Patent data measure the accuracy of assignment of patents to unique individuals. Results are shown for comparison with both the Leuven level 1 (L1) and hand-matched level 2 (L2) datasets. Countries missing precision and recall data had no corresponding Leuven Level 2 IDs in the dataset.

Figure 3 and Table 4 illustrate how the count of patents assigned to unique `dedupe` IDs corresponds to the count assigned to the matching ID from the Leuven dataset. Again, we see that the correlation between patent counts is very high. Implicitly, the ratio of patent counts converges to approximately one for most inventors.

Country	Pearson	Spearman
AT	0.94	0.94
BE	0.93	0.94
CY	0.99	0.96
CZ	0.96	0.96
DE	0.99	0.94
DK	0.87	0.94
EE	0.98	0.96
ES	0.96	0.92
FI	1.00	0.93
FR	0.90	0.94
GB	0.41	0.94
GR	0.90	0.92
HU	0.80	0.91
IE	0.39	0.92
IT	0.87	0.96
LT	0.94	0.95
LU	0.96	0.95
LV	0.96	0.87
MT	1.00	0.96
NL	1.00	0.95
PL	0.83	0.97
PT	0.94	0.90
RO	0.94	0.89
SE	0.97	0.93
SI	0.98	0.94
SK	0.94	0.92

Table 4: Correlation between patent counts assigned to matching Leuven and Dedupe IDs.

6 Discussion

These results point to the potential for rapid, scaleable disambiguation of large patent databases with relatively little human input. Improved data quality or completeness would help improve on these results. In particular, address coverage remains highly variable. At a minimum, however, the techniques demonstrated here appear a natural complement to the extensive data cleaning effort undertaken by Leuven, which may permit even higher levels of precision and recall and eliminate the need for ongoing and intensive human coding.

Nevertheless, data quality continues to pose substantial barriers to progress. The lack of address in particular deprives analysts of a valuable means of disambiguating individuals with common names. Improving address data coverage is obviously beyond the purview of the OECD. Instead, improvements will rely on national patent offices making concerted efforts

A Data attributes and discussion

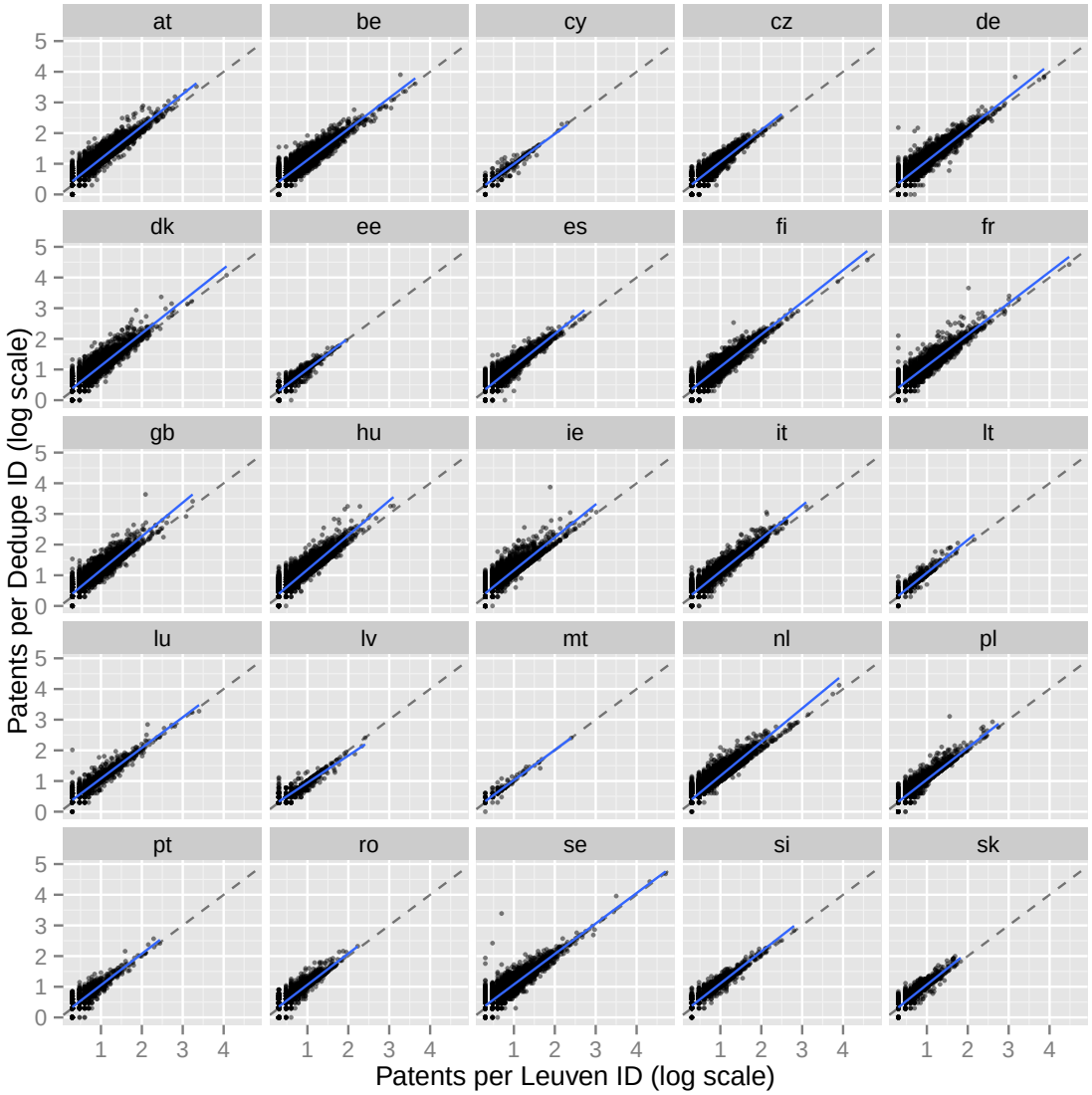


Figure 3: Comparison of patent counts assigned to matching `dedupe` and Leuven unique IDs. Each point represents the number of patents assigned to a unique `dedupe` individual that was also assigned to unique Leuven ID. Points were generated by sampling 10,000 points from all IDs representing 2 or more patents. For IDs that clumped or split their comparison ID, the maximum number of commonly-assigned patents was used.

Country	Pct reduction in unique individuals
AT	68.85
BE	64.69
CY	38.37
CZ	24.12
DE	57.34
DK	60.90
EE	15.53
ES	40.36
FI	56.76
FR	65.51
GB	62.34
GR	50.34
HU	56.32
IE	70.62
IT	63.10
LT	22.37
LU	58.55
LV	19.24
MT	46.69
NL	61.81
PL	21.72
PT	31.57
RO	23.01
SE	55.03
SI	44.68
SK	25.99

Table 5: Percentage reduction in unique IDs by country.

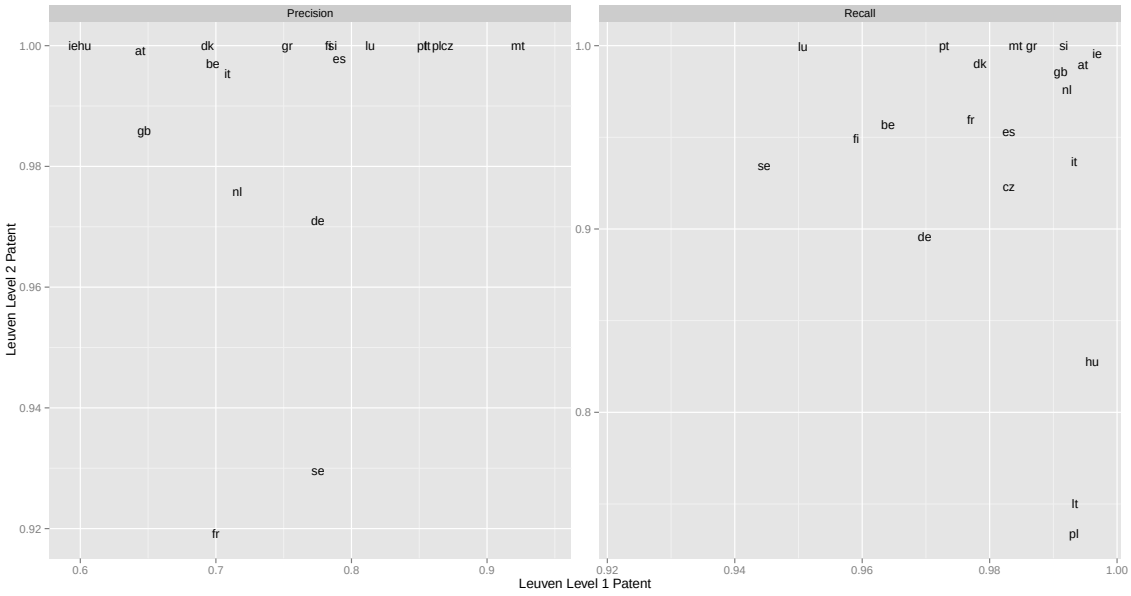


Figure 4: Comparison of precision and recall performance for `dedupe` on Leuven Level 1 (machine-matched) and Level 2 (hand-matched) records. Precision and recall values are computed for the patents assigned to unique Dedupe and Leuven IDs. Countries missing precision and recall data had no corresponding Leuven Level 2 IDs in the dataset.

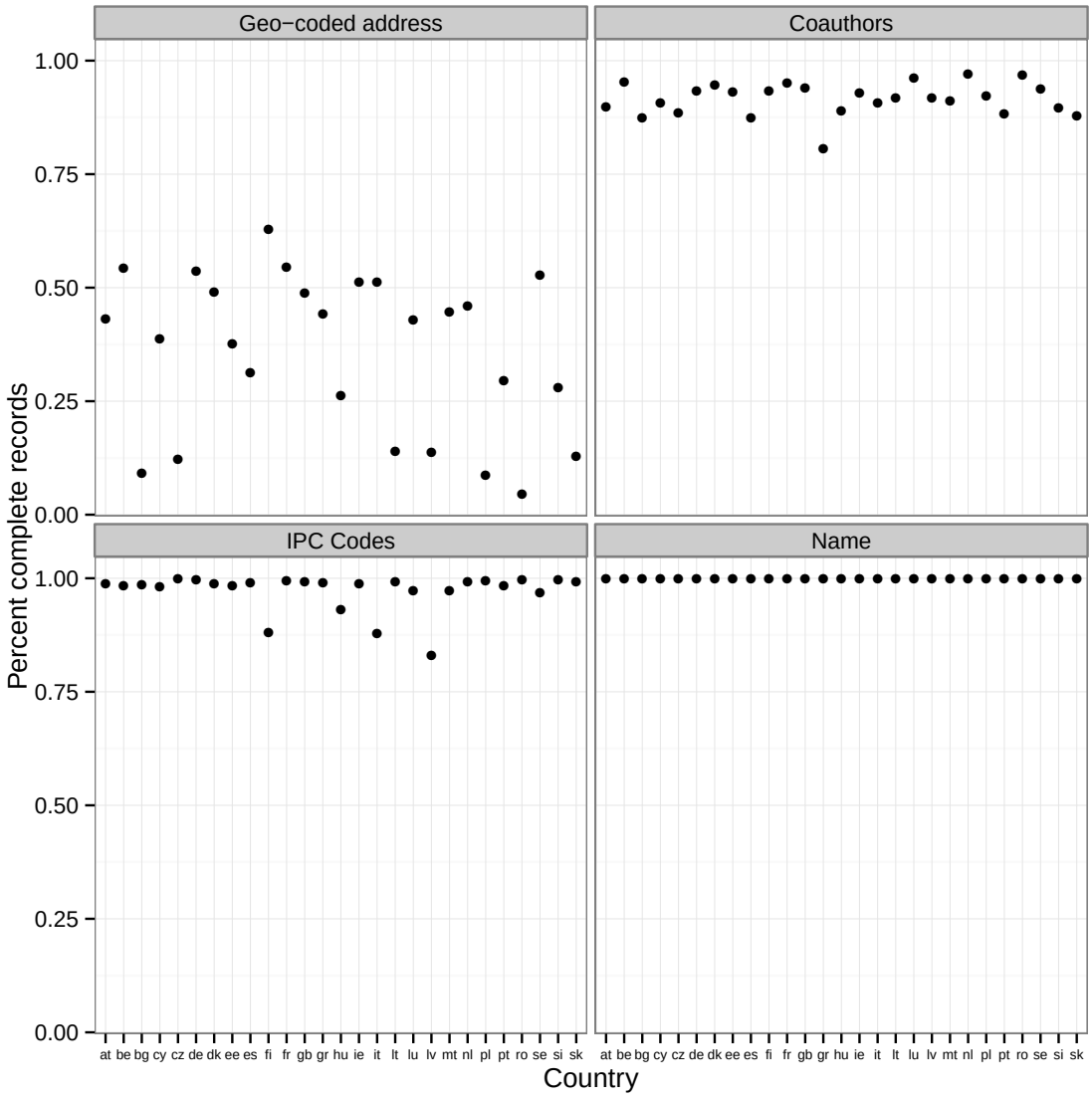


Figure 5: Record completeness for each EU-27 country. This figure illustrates the percent complete records for each field in the `dedupe` person profile input data.

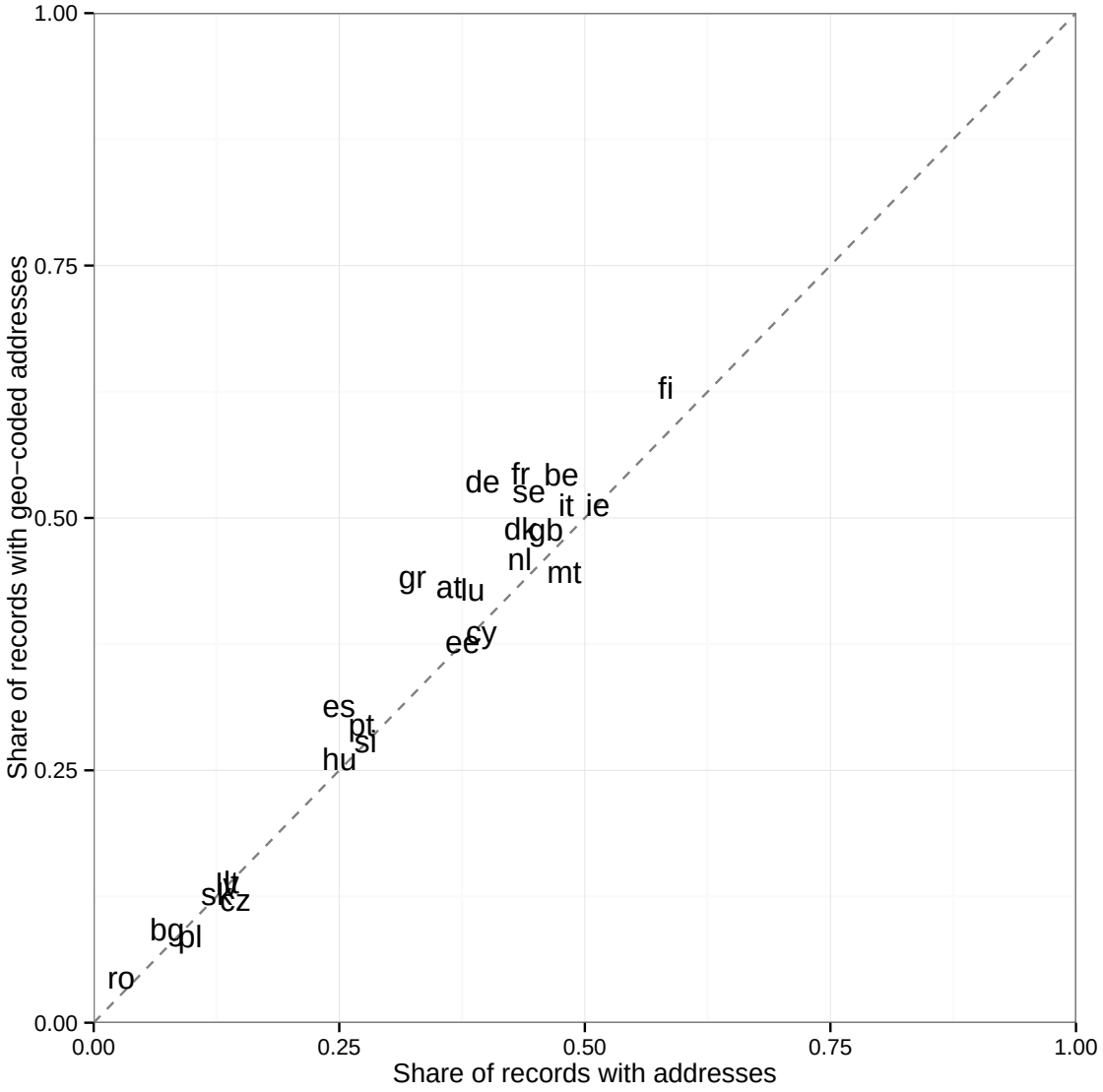


Figure 6: Geo-coding completeness for PATSTAT addresses. This figure compares the share of person records with successfully geo-coded addresses to the presence of addresses in the raw data. In many cases, address data were ambiguous and could not be geocoded. In some cases, however, looking for addresses in the name field improved the overall coverage of geographic data.

B Geocoding logic

The PATSTAT database presents four geocoding challenges: first, address information are inconsistently formatted and vary widely in completeness. Second, some address data are included in the person or company name, rather than in the address field. Third, for European names in particular, names themselves may contain geographic references that do not actually represent the individual's location. For instance, last name variants on the construct "from + place", as in "van/von/de/le/etc + `location`" may encode the place of birth of a distant ancestor, while the person themselves now lives quite far away. Hence in examining names for address data, we must discriminate between true addresses and names with geographic referents. Fourth, some individuals with address information in their name clearly come from countries other than the one in which the patent was filed. For instance, a person living in Germany may be listed on a Dutch patent. We implement algorithms to deal with each problem.

B.1 Fuzzy geocoding

Many open geocoding algorithms proceed by looking for exact matches for a known geographic entity in the address string.[7] This is sub-optimal for PATSTAT given the known problems with address spelling and transcription. We implement the following logic to geocode PATSTAT addresses at the city level:

1. For each address:

    (a) Split the address on spaces

    (b) Identify all unique leading characters in the split address

[7] For instance, the excellent Data Science Toolkit (`http://datasciencetoolkit.org`) looks for exact word matches. It often does not recognize PATSTAT addresses.

(c) Restrict the set of possible cities for comparison to those with first letters in the unique set

2. For each city in the geographic database subset:

   (a) Construct word n-grams from the split address of length equal to the city n-gram (e.g. for "Chicago", use unigrams; for "New York", bigrams, etc)

   (b) Compute the Levenshtein ratio between the city and each n-gram

   (c) Return the nearest-neighbor that exceeds some threshold τ

3. Sort the set of all best matches by Levenshtein ratio, breaking ties by (in order) the ending index of the match in the address string (closer to the end is better); and city population (higher is better)

4. Identify the best city match of all possible matches

5. Return the latitude and longitude for that city

This algorithm is designed to:

- Allow for fuzzy matches wherein the city name is misspelled
- Prefer matches later in the string, on the assumption that cities are listed towards the end of addresses
- Resolve any remaining ambiguity in favor of higher-population cities
- Speed geo-coding by restricting the comparison set of cities to likely matches based on leading letters
- Avoid mis-identifying cities by looking at the relevant n-gram, rather than the entire string (e.g., to avoid coding "Frankfurt am Main" as "Mainz" due to the similarity of "Main" and "Mainz")

The full implementation of the algorithm is available as the `fuzzygeo` module for python, at https://github.com/markhuberty/fuzzygeo. Runtimes scale with the number of possible comparison cities. For medium-sized countries, this equates to anywhere from 25-50ms per address. This can go as high as 100-200ms for a large country with many cities, like the United States. Runtimes can be improved by sub-setting cities using the first two letters of all address ngrams, at obvious risk to recall.

B.2 Name parsing

We implement the following logic to parse address information from names:

1. Split the name at most once on a sequence of numbers bound by spaces on either side

2. If the split returns two components, check the second component as follows:

   (a) Check for a potential two-letter country code at the end of the string. If that country is not found, or is the same as the country of the patent filing, geo-code for that country; else geo-code for the alternate country

   (b) Remove the country code, if found, from the address component

   (c) Geo-code the address component using the appropriate fuzzy geo-coding algorithm as described in B.1

This algorithm assumes:

- That names containing addresses will have a street number dividing the name from the address

- That individuals may originate in countries other than that in which the patent is filed

This algorithm will miss geographic content in names that have no street code. For instance "John Smith Chicago" will not return a valid latitude / longitude, while "John Smith 1234 N. Clark St. Chicago" will. This is an unavoidable consequence of trying to explicitly avoid false matches in names that, for historic reasons, embed geographic data.

C Sample name output

This section provides tables with sample ID matching between the `dedupe` results and their corresponding Leuven IDs. Tables for each country in the EU-25 (excluding Cyprus and Malta) are provided. "Clumping" tables indicate where a single `dedupe` ID consolidated multiple Leuven IDs. "Splitting" tables indicate the reverse: where names consolidated to a single Leuven ID were split across multiple `dedupe` IDs.

C.1 Leuven IDs split by `dedupe`

Dedupe ID	Leuven ID	Name	Country
65065	37599889	WARTHNER, Hubert	AT
65066	37599889	WARTHNER, Hubert	AT
65067	37599889	WA1/4RTHNER, Hubert	AT
65068	37599889	WA1/4RTHNER, Hubert	AT
65069	37599889	WA1/4RTHNER, Hubert	AT
23632	14473277	* INTERUNIVERSITAIR MICRO-ELEKTRONICA	BE
23632	14473277	IMEC	BE
23632	14473277	Imec	BE
23632	14473277	IMEC Co.	BE
23632	14473277	IMEC CORP.	BE
59039	10147956	* G. D. SOCIETA PER AZIONI	IT
59040	10147956	* G. D SOCIETA' PER AZIONI	IT
59039	10147956	* G. D SOCIETA PER AZIONI	IT

59040	10147956	* G D SOCIETA' PER AZIONI	IT
59040	10147956	* G D SOCIETA 'PER AZIONI	IT
130988	34019163	* THE BOC GROUP PLC	GB
159982	34019163	BOC GROUP P L C	GB
159983	34019163	BOC GROUP PLC	GB
159983	34019163	BOC Group plc	GB
159984	34019163	BOC GROUP, P.L.C. (THE)	GB
340950	26840279	PHILIPPE, MICHEL	FR
340951	26840279	PHILIPPE, MICHEL	FR
340952	26840279	PHILIPPE, Michel	FR
340953	26840279	PHILIPPE, Michel	FR
340954	26840279	PHILIPPE, Michel	FR
895397	23922940	MALLER, Thomas	DE
895398	23922940	MALLER, Thomas	DE
895399	23922940	MALLER, Thomas	DE
895400	23922940	MALLER, Thomas	DE
895401	23922940	MALLER, Thomas	DE
8589	36143041	VUTCH - CHEMITEX, SPOL. S R.O.	SK
8590	36143041	VUTCH-CHEMITEX	SK
8591	36143041	VUTCH-CHEMITEX, A.S.	SK
8566	36143041	VUTCH-CHEMITEX, SPOL. S R. O.	SK
8592	36143041	VUTCH-CHEMITEX SPOL. S R. O.	SK
47482	33865644	* TELEFONAKTIEBOLAGET L M ERICSSON	SE
47483	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON	SE
47482	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON	SE
47482	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON PUBL	SE
48248	33865644	ERICSSON	SE
10494	33829731	TECNIMEDE - SOCIEDADE TECNICO-MEDICINAL, S.A.	PT
10495	33829731	TECNIMEDE SOCIEDADE TECNICO MEDICINAL S.A.	PT
10496	33829731	TECNIMEDE- SOCIEDADE TECNICO-MEDICINAL, S.A.	PT
10496	33829731	TECNIMEDE SOCIEDADE TECNICO-MEDICINAL, S.A.	PT

10497	33829731	TECNIMEDE-SOCIEDADE TECNICO-MEDICINAL, S.A.	PT
7651	173039	A B B SPO&LSTROK;KA Z OGRANICZON&AOGON; ODPOWI...	PL
7652	173039	A B B SP.Z OO.	PL
7681	173039	ABB OY	PL
7682	173039	ABB SP. Z O. O.	PL
7682	173039	ABB SP Z. O. O.	PL
35210	75994	3D HISTECH KFT	HU
35211	75994	3D HISTECH KFT.	HU
35212	75994	3D HISTECH KFT.	HU
35214	75994	3DHISTECH KFT	HU
35214	75994	3DHISTECH KFT.	HU
9245	3746961	BRONWAY RESEARCH LIMITED	IE
9246	3746961	BRONWAY RESEARCH LIMITED	IE
9247	3746961	BRONWAY RESEARCH LIMITED	IE
9248	3746961	Bronway Research Limited	IE
9247	3746961	Bronway Research Limited	IE
6589	33731928	TARTU UELIKOOL	EE
6590	33731928	TARTU UELIKOOL	EE
6592	33731928	TARTU ULIKOOL	EE
6593	33731928	Tartu Alikool	EE
101587	21758654	MARTINEZ MARTINEZ, ANTONIO	ES
101588	21758654	MARTINEZ MARTINEZ, Antonio	ES
101589	21758654	MARTINEZ MARTINEZ, Antonio	ES
101590	21758654	MARTINEZ MARTINEZ, Antonio	ES
101591	21758654	MARTINEZ MARTINEZ, Antonio	ES
1300	35269555	UNIBIND (CYPRUS) LIMITED	CY
1301	35269555	UNIBIND (CYPRUS) LIMITED	CY
1300	35269555	UNIBIND [CYPRUS] LIMITED	CY
1300	35269555	UniBind (Cyprus) Limited	CY
1300	35269555	Unibind (Cyprus) Limited	CY
10301	4657566	CESKE VYSOKE UCENI TECHNICKE V. PRAZE	CZ

10302	4657566	CESKE VYSOKE UCENI TECHNICKE V PRAZE	CZ
10290	4657566	CESKE VYSOKE UCENI TECHNICKE V PRAZE	CZ
10290	4657566	Ceske Vysoke Uceni Technicke V Praze	CZ
10303	4657566	Ceske vysoke uceni technicke v Praze	CZ
49164	26842439	* KONINKLIJKE PHILIPS ELECTRONICS N.V.	NL
49164	26842439	KONINKLIJKE PHILIPS ELECTRONICS N.V.	NL
49164	26842439	* N V PHILIPS' GLOEILAMPENFABRIEKEN	NL
49164	26842439	* N V PHILIPS‘ GLOEILAMPENFABRIEKEN	NL
49164	26842439	* N.V. PHILIPS' GLOEILAMPENFABRIEKEN	NL
2192	14643060	* INSTITUT JOZEF STEFAN	SI
2192	14643060	INSTITUT JOZEF STEFAN	SI
2192	14643060	INSTITUT " JOZEF STEFAN"	SI
2192	14643060	INSTITUT JOZEF STEFAN	SI
2192	14643060	INSTITUT 'JOZEF STEFAN'	SI
14814	14660477	INSTITUTUL POLITEHNIC TRAIAN VUIA, TIMISOARA	RO
14813	14660477	INSTITUTUL POLITEHNIC TRAIAN VUIA, TIMISOARA	RO
14805	14660477	INSTITUTUL POLITEHNIC TRAIAN VUIA, TIMISOARA,	RO
14815	14660477	INSTITUTUL POLITEHNIC TRAIAN VUIA, TIMISOARA	RO
14816	14660477	INSTITUTUL POLITEHNIC TRAIAN VUIA,TIMISOARA	RO
40820	18084510	KOBENHAVNS UNIVERSITET	DK
40821	18084510	KOBENHAVNS UNIVERSITET	DK
40822	18084510	KOBENHAVNS UNIVERSITET	DK
40823	18084510	Kobenhavns Universitet	DK
40817	18084510	Kobenhavns Universitet	DK
774	2083626	BARAUSKAS, ARVYDAS	LT
776	2083626	BARAUSKAS,ARVYDAS	LT
1808	8695192	* EURO-CELTIQUE S.A	LU
1809	8695192	* EUROCELTIQUE S.A	LU
1809	8695192	* EUROCELTIQUE S.A.	LU
2903	8695192	EURO - CELTIQUE, S/A	LU
2903	8695192	EURO - CELTIQUE S.A.	LU

2302	16593879	KALVINS, IVARS	LV
2303	16593879	KALVINS, Ivars	LV
2304	16593879	KALVINS, Ivars	LV
2305	16593879	KALVINS IVARS	LV
2304	16593879	Kalvins, Ivars	LV
162	1464780	ART PRODUCTIONS LIMITED	MT
163	1464780	Art Productions Limited	MT
163	1464780	ART PRODUCTIONS LTD.	MT
31447	24869999	* NOKIA CORPORATION	FI
31447	24869999	* NOKIA OY	FI
31447	24869999	* NOKIA OYJ	FI
31447	24869999	NOKIA	FI
31447	24869999	Nokia	FI
1989	1389555	ARISTOTLE UNIVERSITY OF THESSALONIKI- Researc...	GR
1988	1389555	ARISTOTLE UNIVERSITY OF THESSALONIKI- Researc...	GR
1987	1389555	ARISTOTLE UNIVERSITY OF THESSALONIKI-RESEARCH ...	GR

C.2 Leuven IDs clumped by dedupe

Dedupe ID	Leuven ID	Name	Country
25642	2721980	BERNER FRANTS	AT
25642	2722272	BERNER, FRANZ	AT
25642	2722272	BERNER, Franz	AT
25642	2721981	BERNER FRANZ	AT
25642	2721982	BERNER FRANZ DIPL. ING.	AT
25684	1891734	BAETEN, ROGER, SEPTESTRAAT 27 B-2640 MORTSEL, BE	BE
25684	2241889	BASTIAENS, LUC	BE
25684	2241889	BASTIAENS, Luc	BE
25684	2241889	Bastiaens, Luc	BE
25684	2241896	Bastiaens, Luc, c/o Agfa-Gevaert N.V.	BE
58930	4451152	CARMINATI PAOLO	IT

58930	31504733	* SIGMA-TAU INDUSTRIE FARMACEUTICHE RIUNITE S P A	IT
58930	31504733	* SIGMA-TAU INDUSTRIE FARMACEUTICHE RIUNITE SPA	IT
58930	2265823	BATTISTINI	IT
58930	2265826	BATTISTINI, ALBERTO	IT
118898	1918877	* BAILEY	GB
118898	1919210	* BAILEY CONCEPTS LTD	GB
118898	2962441	* BIOTICA TECHNOLOGY LIMITED	GB
118898	3678973	* BRIGGS	GB
118898	3679363	* BRIGGS IRRIGATION	GB
133280	5673120	* COLAS S.A.	FR
133280	35798536	* VERNET S.A.	FR
133280	205881	ABELARD, Franck, c/o THOMSON multimedia	FR
133280	205882	ABELARD, Franck, Thomson multimedia	FR
133280	205880	ABELARD, Franck Thomson multimedia	FR
480300	7562763	DR. SCHNEIDER, WERNER	DE
480300	7581600	DR.SCHNEIDER, WERNER	DE
480300	7581600	Dr.Schneider, Werner	DE
480300	30249877	Schneider, Jens	DE
480300	30246502	SCHNEIDER + NOELKE GMBH	DE
4422	18110863	KOCIS, DUSAN	SK
4422	18110894	KOCIS, Dusan	SK
4422	18110863	KOCIS DUSAN	SK
4422	18110863	Kocis, Dusan	SK
4422	18110866	KOCIS DUSAN,KOCIS IVAN	SK
48248	160224	AASE, Karin	SE
48248	160175	AASE KARIN	SE
48248	160224	Aase, Karin	SE
48248	160229	AASE, Karin, c/o The Ludwig Inst.Cancer Research	SE
48248	160230	AASE, Karin, Ludwig Institute for Cancer Research	SE
2674	1336169	ARAUJO SOARES DA SILVA, PATRICIO, MANUEL, VIEIRA	PT
2674	1336169	ARAUJO SOARES DA SILVA, PATRICIO MANUEL VIEIRA	PT

2674	1336172	ARAUJO SOARES DA SILVA, PatrAcio Manuel Vieira	PT
2674	1336167	ARAUJO SOARES DA SILVA PATRICIO MANUEL VIEIRA	PT
2674	6254377	DA SILVA, PATRICIO MANUEL VIEIRA ARAUJO SOARES	PT
49847	27496932	PRZEDSIA(r)BIORSTWO APLIKACJI INAY=YNIERSKICH ...	PL
49847	27496934	PRZEDSIA(r)BIORSTWO BUDOWY SZYBOW SPOEKA AKCYJNA	PL
49847	27496935	PRZEDSIA(r)BIORSTWO INTERMAG SPOEKA Z O.O.	PL
49847	27496936	PRZEDSIA(r)BIORSTWO PRODUKCYJNO-HANDLOWE ARMAT...	PL
49847	27496937	PRZEDSIA(r)BIORSTWO PRODUKCYJNO-HANDLOWO-USEUG...	PL
10145	1805937	B. KOVACS, ATTILA	HU
10145	1798755	B. KOVACS ATTILA	HU
10145	1798756	B. KOVACS,ATTILA	HU
10145	1798757	B. KOVACZ, ATTILA	HU
10145	2190617	BARTHA LASLO	HU
7280	6081307	* CRONIN BUCKLEY STEEL ERECTORS LIMITED	IE
7280	20042555	* LETT RESEARCH AND DEVELOPMENT LIMITED	IE
7280	34948696	* TSUNAMI PHOTONICS LIMITED	IE
7280	451559	Ahs, David c/o Microsoft (EPDC)	IE
7280	451561	Ahs, David, Microsoft, European Product Devel...	IE
1527	2504679	BELLAKEM OE	EE
1527	2504680	BELLAKEM OJU	EE
1527	2504675	BELLAKEM OU	EE
1527	2504675	BELLAKEM OUE	EE
1527	2504675	BELLAKEM OUE,	EE
30924	5814283	* CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS	ES
30924	5787303	Conejo Superior de Investigaciones Cientificas	ES
30924	5813966	Consego Superior de Investigaciones Cientificas	ES
30924	5814243	CONSEJO SUPEIOR DE INVESTIGACIONES CIENTIFICAS	ES
30924	5814247	CONSEJO SUPERIOR	ES
500	5947196	Costa, Costas	CY
500	5947200	COSTA, COSTAS, N.	CY
500	5947197	COSTA, COSTAS N.	CY

500	5947197	COSTA, Costas N.	CY
500	5945899	COSTA COSTAS N.	CY
18282	14628711	INST OF EXPERIMENTAL BOTANY, ACADEMY OF SCIENC...	CZ
18282	14632943	INST. OF ORGANIC CHEMISTRY AND BIOCHEMISTRY OF...	CZ
18282	14652121	Institute of Experimental Bontany of the Acade...	CZ
18282	14652139	Institute of Experimental Botany, Academy of ...	CZ
18282	14652139	INSTITUTE OF EXPERIMENTAL BOTANY, ACADEMY OF S...	CZ
57725	2399348	BECKERS, LUCAS J.A.M., C/O INT. OCTROOIBUREAU ...	NL
57725	3104536	BLOM, GERARD, C/O INT. OCTROOIBUREAU B.V., NL-...	NL
57725	3202570	BOEZEN, HENDRIK, C/O INT. OCTROOIBUREAU B.V, N...	NL
57725	3266479	BOLT, JACOB HENDRIK, C/O INT. OCTROOIBUREAU B....	NL
57725	3735744	BROER, DIRK JAN, C/O INT. OCTROOIBUREAU B.V., ...	NL
3171	6236152	.D. LEK, TOVARNA FARMACEVTSKIH IN KEMICNIH IZD...	SI
3171	19906684	LEK, TOVARNA FARMACEUTSKIH IN KEMICNIH IZDELKO...	SI
3171	19906690	LEK, TOVARNA FARMACEVTSKIH	SI
3171	19906690	LEK, Tovarna Farmacevtskih	SI
3171	19906690	LEK, tovarna farmacevtskih	SI
14326	14656948	INSTITUTUL DE CERCETARE PENTRU RAFINARII A!I P...	RO
14326	14658195	INSTITUTUL DE CERCETARI PENTRU EAFINARII SI PE...	RO
14326	14658357	INSTITUTUL DE CERCETARI PENTRU RAFDINARII SI P...	RO
14326	14658396	INSTITUTUL DE CERCETARI PENTRU RAFINARII A!I P...	RO
14326	14658365	INSTITUTUL DE CERCETARI PENTRU RAFINARII SI PE...	RO
18144	24595054	Nielsen,John Godsk	DK
18144	108794	A. EDVARDSEN	DK
18144	992291	ANDERSEN, ANDRES	DK
18144	992337	ANDERSEN, ARTHUR	DK
18144	992387	ANDERSEN, B GE	DK
1325	11364568	GRAZULEVICIUS JUOZAS	LT
1325	11364577	GRAZULEVICIUS, JUOZAS V.	LT
1325	11364571	GRAZULEVICIUS JUOZAS V.	LT
1325	11364577	Grazulevicius, Juozas V.	LT

1325	11364581	GRAZULEVICIUS, JUOZAS VIDAS	LT
1811	8699148	* EUROPEAN COMMUNITY	LU
1811	8699275	* EUROPEAN ECONOMIC COMMUNITY REPRESENTED BY T...	LU
1811	8693510	EURATOM	LU
1811	8693510	Euratom	LU
1811	8693531	EURATOM (EUROPAEISCHE ATOMGEMEINSCHAFT)	LU
2301	16593877	KALVINS, I., Latvian Inst. of Organic Synthesis	LV
2301	16593878	KALVINS, I., Latvian Inst. Oraganic Synthesis	LV
2301	16593901	KALVINS, Ivars; Latvian Inst. of Organic Synth...	LV
2301	16593897	KALVINS, Ivars Latvian Inst. of Organic Synthesis	LV
2301	16593900	KALVINS, Ivars, Latvian Institute of Organic S...	LV
423	27524848	PULE', JOSEPH	MT
423	27524848	PULE', Joseph	MT
423	27524853	PULE, JOSEPH	MT
423	27524853	PULE, Joseph	MT
423	27524844	PULE JOSEPH	MT
41302	10313226	GARCIA MARTIN, MIGUEL	FI
41302	10313227	GARCIA MARTIN, MIGUEL ANGEL	FI
41302	10313228	GARCIA MARTIN, Miguel, Angel	FI
41302	10318203	GARCIA, MIGUEL	FI
41302	10318203	GARCIA, Miguel	FI
1842	967164	ANAGNOSTOPOULOS, A., PANAGIOTIS	GR
1842	967167	ANAGNOSTOPOULOS, ANTONIOS	GR
1842	967167	ANAGNOSTOPOULOS ANTONIOS	GR
1842	967167	Anagnostopoulos, Antonios	GR
1842	967103	ANAGNOSTOPOULOS, ANTONIOS P.	GR

C.3 Leuven Level 2 IDs split by dedupe

Dedupe ID	Leuven ID	Name	Country
22970	9636600	* FRANZ PLASSER BAHNBAUMASCHINEN INDUSTRIEGESE...	AT

22970	9636600	* FRANZ PLASSER BAHNBAUMASCHINEN INDUSTRIEGESE...	AT
22970	9636600	* FRANZ PLASSER BAHNBAUMASCHINEN-INDUSTRIEGESE...	AT
22970	9636600	* FRANZ PLASSER BAHNBAUMASCHINEN-INDUSTRIEGESE...	AT
22970	9636600	* FRANZ PLASSER BAHNBAUMASCHINEN-INDUSTRIEGESE...	AT
23632	14473277	* INTERUNIVERSITAIR MICRO-ELEKTRONICA	BE
23632	14473277	IMEC	BE
23632	14473277	Imec	BE
23632	14473277	IMEC Co.	BE
23632	14473277	IMEC CORP.	BE
59039	10147956	* G. D. SOCIETA PER AZIONI	IT
59040	10147956	* G. D SOCIETA' PER AZIONI	IT
59039	10147956	* G. D SOCIETA PER AZIONI	IT
59040	10147956	* G D SOCIETA' PER AZIONI	IT
59040	10147956	* G D SOCIETA 'PER AZIONI	IT
130988	34019163	* THE BOC GROUP PLC	GB
159982	34019163	BOC GROUP P L C	GB
159983	34019163	BOC GROUP PLC	GB
159983	34019163	BOC Group plc	GB
159984	34019163	BOC GROUP, P.L.C. (THE)	GB
133284	5751320	* COMMISSARIAT A L 'ENGERGIE ATOMIQUE	FR
133284	5751320	* COMMISSARIAT A L'ENERGIE ATOMIQUE	FR
179362	5751320	C.E.A.	FR
179363	5751320	CEA	FR
179364	5751320	CEA	FR
328189	31483378	* SIEMENS AG	DE
328189	31483378	* SIEMENS AKTIENGESELLSCHAFT	DE
328189	31483378	* SIEMENS AKTIENGESELLSHAFT	DE
336341	31483378	Aktiengesellschaft; Siemens	DE
336341	31483378	Aktiengesellschaft,Siemens	DE
47482	33865644	* TELEFONAKTIEBOLAGET L M ERICSSON	SE
47483	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON	SE

47482	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON	SE
47482	33865644	* TELEFONAKTIEBOLAGET LM ERICSSON PUBL	SE
48248	33865644	ERICSSON	SE
3790	6250602	DA COSTA LAGE, ANTONIO MANUEL	PT
3790	6250602	Da Costa Lage, Antonio Manuel	PT
7651	173039	A B B SPO&LSTROK;KA Z OGRANICZON&AOGON; ODPOWI...	PL
7652	173039	A B B SP.Z OO.	PL
7681	173039	ABB OY	PL
7682	173039	ABB SP. Z O. O.	PL
7682	173039	ABB SP Z. O. O.	PL
25128	22491374	MELYEPITESI TERVEZO VALLALAT	HU
25129	22491374	MELYEPITESI TERVEZOE VALLALAT	HU
25130	22491374	MELYEPITESI TERVEZOE VALLALAT,HU	HU
8940	3404519	BOSTON SCIENT LTD.	IE
8939	3404519	BOSTON SCIENTIFIC LIMITED	IE
8939	3404519	Boston Scientific Limited	IE
32652	478543	AIRBUS ESPA I A, S.L.	ES
32653	478543	AIRBUS ESPA I A S.L.	ES
32654	478543	AIRBUS ESPAA+-A, S.L.	ES
32654	478543	AIRBUS ESPAA+-A S.L.	ES
32654	478543	AIRBUS ESPANA , S.L.	ES
26091	21676388	MARS A. S.	CZ
26092	21676388	MARS A.S.	CZ
26092	21676388	MARS, S.R.O.	CZ
49164	26842439	* KONINKLIJKE PHILIPS ELECTRONICS N.V.	NL
49164	26842439	KONINKLIJKE PHILIPS ELECTRONICS N.V.	NL
49164	26842439	* N V PHILIPS' GLOEILAMPENFABRIEKEN	NL
49164	26842439	* N V PHILIPS‘ GLOEILAMPENFABRIEKEN	NL
49164	26842439	* N.V. PHILIPS' GLOEILAMPENFABRIEKEN	NL
2210	173039	ABB d.o.o.	SI
18147	24996280	* NOVO NORDISK A/S	DK

18147	24996280	A/S NOVO NORDISK	DK
49453	24996280	NOVA NORDISK A/S	DK
49453	24996280	Nova Nordisk A/S	DK
18147	24996280	Novo Nordisk A/S	DK
1398	13270966	HILTI AG	LT
1399	13270966	Hilti Aktiengesellschaft	LT
1821	26358567	* PAUL WURTH S.A	LU
1821	26358567	* PAUL WURTH S.A.	LU
1821	26358567	* PAUL WURTH SA	LU
1821	26358567	PAUL WURTH	LU
1821	26358567	Paul Wurth	LU
474	32634657	STMicroelectronics Ltd	MT
474	32634657	STMicroelectronics Ltd.	MT
31447	24869999	* NOKIA CORPORATION	FI
31447	24869999	* NOKIA OY	FI
31447	24869999	* NOKIA OYJ	FI
31447	24869999	NOKIA	FI
31447	24869999	Nokia	FI

C.4 Leuven Level 2 IDs clumped by dedupe

Dedupe ID	Leuven ID	Name	Country
23017	24986772	NOVARTIS AG	AT
23017	24986772	NOVARTIS GMBH	AT
23017	24987227	NOVARTIS PHARMA AG	AT
23017	24987227	Novartis Pharma AG	AT
23017	24987227	NOVARTIS PHARMA GMBH	AT
23632	14473277	* INTERUNIVERSITAIR MICRO-ELEKTRONICA	BE
23632	14473277	IMEC	BE
23632	14473277	Imec	BE
23632	14473277	IMEC Co.	BE

23632	14473277	IMEC CORP.	BE
75331	3187124	Boehringer Ingelheim Italia	IT
75331	3187124	BOEHRINGER INGELHEIM ITALIA S. P. A.	IT
75331	3187124	BOEHRINGER INGELHEIM ITALIA S P A	IT
75331	3187124	BOEHRINGER INGELHEIM ITALIA S.P.A	IT
75331	3187124	BOEHRINGER INGELHEIM ITALIA S.P.A.	IT
118898	7909433	* KODAK LIMITED	GB
118898	26795070	* PFIZER LIMITED	GB
118898	15987300	JOHNSON & JOHNSON	GB
118898	15987300	JOHNSON & JOHNSON LIMITED	GB
118898	15987300	Johnson & Johnson Limited	GB
133268	4631169	* CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	FR
133268	4631169	(CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE...	FR
133268	4631169	) CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE...	FR
133268	4631169	- CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE...	FR
133268	14645395	INSERM	FR
326983	6308273	* DAIMLER BENZ AKTIENGESELLSCHAFT	DE
326983	6308273	* DAIMLER-BENZ A G	DE
326983	6308273	* DAIMLER-BENZ AG	DE
326983	6308273	* DAIMLER-BENZ AKTIENGESELLSCHAFT	DE
326983	6308670	* DAIMLERCHRYSLER AG	DE
47213	1606297	* AB ASTRA	SE
47213	1606297	* ASTRA AKTIEBOLAG	SE
47213	1606297	* ASTRA AKTIEBOLAGET	SE
47213	1608007	* ASTRAZENECA AB	SE
47213	1606297	AB, ASTRA	SE
3790	6250602	DA COSTA LAGE, ANTONIO MANUEL	PT
3790	6250602	Da Costa Lage, Antonio Manuel	PT
7651	173039	A B B SPO&LSTROK;KA Z OGRANICZON&AOGON; ODPOWI...	PL
9920	1292153	APPLIED MATERIAL CO., LTD.	HU
7136	968037	* ANALOG DEVICES B.V.	IE

7136	968037	ANALOG DEVICES, B.V.	IE
7136	968037	ANALOG DEVICES B.V.	IE
7136	968037	ANALOG DEVICES BV	IE
7136	968037	Analog Devices, B.V.	IE
50844	4517266	Casanova PA(c)rez, Elena Maria	ES
50844	4517268	Casanova PA(c)rez, Gregorio	ES
50844	4517271	Casanova PA(c)rez, Maria Mercedes	ES
6316	173039	ABB S.R.O.	CZ
49237	30169087	* SCHLUMBERGER LIMITED	NL
49237	30170026	* SCHLUMBERGER TECHNOLOGY BV	NL
49237	30169087	SCHLUMBERGER CA LTD.	NL
49237	30169087	SCHLUMBERGER CANADA LIMITED	NL
49237	30169424	SCHLUMBERGER HOLDINGS	NL
2210	173039	ABB d.o.o.	SI
18068	5727531	* COLOPLAST A/S	DK
18068	5727531	Coloplast A	DK
1398	13270966	HILTI AG	LT
1821	26358567	* PAUL WURTH S.A	LU
1821	26358567	* PAUL WURTH S.A.	LU
1821	26358567	* PAUL WURTH SA	LU
1821	26358567	PAUL WURTH	LU
1821	26358567	Paul Wurth	LU
474	32634657	STMicroelectronics Ltd	MT
474	32634657	STMicroelectronics Ltd.	MT
31360	173039	* ABB OY	FI
31360	173039	ABB AB	FI
31360	173039	ABB CO.	FI
31360	173039	ABB CORP.	FI
31360	173039	ABB OY	FI
3158	13470090	HOECHST AKTIENGESELLSCHAFT	GR

References

Bilenko, M. Y. (2006). *Learnable similarity functions and their application to record linkage and clustering*, volume 67.

Callaert, J., Du Plessis, M., Grouwels, J., Lecocq, C., Magerman, T., Peeters, B., Song, X., Van Looy, B., and Vereyen, C. (2011). Patent statistics at eurostat: Methods for regionalisation, sector allocation and name harmonisation. *Eurostat Methodologies and Working Papers*.

Lai, R., D'Amour, A., Yu, A., Sun, Y., Torvik, V., and Fleming, L. (2011). Disambiguation and co-authorship networks of the us patent inventor database. Technical report, Harvard Institute for Quantitative Social Science, Cambridge, MA.

Raffo, J. and Lhuillery, S. (2009). How to play the "names game": Patent retrieval comparing different heuristics. *Research Policy*, 38(10):1617–1627.

Torvik, V. I., Weeber, M., Swanson, D. R., and Smalheiser, N. R. (2005). A probabilistic similarity metric for medline records: a model for author name disambiguation. *Journal of the American Society for information science and technology*, 56(2):140–158.