U.S. Patent and Trademark Office
OFFICE OF THE CHIEF ECONOMIST
Economic Working Paper Series

Technical Documentation for Matching Patents and Trademarks to the 2017 National Establishment Time Series Database

Ryan Hughes, *Economist*
Charles deGrazia, *Assistant Professor*
Julian Kolev, *Economist*

USPTO Economic Working Paper No. 2021-4
September 2021

UNITED STATES
PATENT AND TRADEMARK OFFICE
uspto

Technical Documentation for Matching Patents and Trademarks to the 2017 National Establishment Time Series Database

September 2021

Ryan Hughes[a,b] Charles deGrazia[c] Julian Kolev[a]

[a]United States Patent and Trademark Office
[b]Addx Corporation
[c]Leonard De Vinci Business School

ABSTRACT: This paper documents the matching procedures used to generate crosswalks between the USPTO's patent and trademark datasets and the 2017 National Establishment Time Series (NETS) database. The matching procedures implement a combination of entity resolution techniques including name and address standardization, heuristic matching, hierarchical disambiguation, and manual verification. Using these methods, the Office of the Chief Economist (OCE) matched 88% of patents granted and 73% of trademarks from 1990 through 2016 to entities in the NETS database. The resulting crosswalks offer a foundation for a wide range of work to investigate the relationship between intellectual property and firm establishments.

JEL: O31, O34, O51, C57, C81
Keywords: patent, trademark, intellectual property, NETS, entity resolution, disambiguation

For more information about the USPTO's Office of the Chief Economist, visit www.uspto.gov.economics.

1. Introduction

Intellectual property and other intangible assets have become increasingly important drivers of business activity across diverse sectors of the economy. In light of this, it is vital for policymakers, academic researchers, and business owners to understand the economic impact of intellectual property activity at the firm, industry, and country levels. To quantify the economic role of patents and trademarks, two prominent forms of intellectual property (IP), one must have access to detailed, firm-level data linking the performance of firms to their patent and trademark holdings. The Office of the Chief Economist (OCE) within the United States Patent and Trademark Office (USPTO) responded to the pressing need for a high-quality dataset of this type by linking the USPTO's intellectual property data on patent and trademark assignees with the National Establishment Time Series (NETS) data provided by Walls & Associates, which is created using annual files from Dun & Bradstreet (D&B).[1] Importantly, the NETS database covers not only publicly-traded firms, but also privately-held corporations, small businesses, new ventures, educational institutions, and non-profit and government organizations. This breadth of coverage provides a comprehensive view of the intellectual property ecosystem, and the datasets described in this paper have the potential to address a wide range of research and policy questions related to the use of patents and trademarks in society.

This document details the inputs, methods, and results of the processes that OCE used to create linkages between the USPTO's IP datasets and the NETS database. Section 2 reviews the information on patents and trademarks that comprise the IP data. Section 3 describes the NETS database. Section 4 introduces the Doherr SearchEngine (DSE), a free software program with powerful entity resolution and record linkage capabilities. Section 5 gives the procedure used for matching patent data to the NETS database. Section 6 presents the results of matching the patents and NETS data. Section 7 explains the steps used to match trademarks with the NETS database. Section 8 provides statistics on the performance of the trademark to NETS matching, and Section 9 concludes.

2. Intellectual property datasets

I. Patents

The OCE used the PatentsView (PV) database of all patents granted by the USPTO from the years 1990 through 2016.[2] We used the following specific files for this purpose: i) *patent.dta*, containing

[1] Due to the proprietary nature of the NETS database, it is not possible to publicly release the crosswalk between NETS and USPTO's intellectual property datasets at this time.
[2] The PatentsView dataset is publicly available at https://patentsview.org/.

basic patent information (including the patent number) on over 6 million patents granted over the period 1990-2016, ii) *rawassignee.dta*, containing the assignee at grant, and iii) *rawlocation.dta*, containing assignee city and state location information. We linked these files together using the patent number (given by the variable *patent_id*), and dropped entries for which the organization variable was missing (that is, patents granted to individuals). We standardized assignee names and city locations using the downloadable Stata packages, *stnd_compname* and *stnd_address*[3] in order to generate a final dataset of patent-assignee observations.[4]

II. Trademarks

For the purpose of matching trademarks to the NETS database, OCE created a dataset consisting of all trademarks registered to US-based firms or organizations in the USPTO's Trademarks Case Files (TMCF) database from 1990 to 2016.[5] From the TMCF, the file *owner.dta* contains basic information on trademark ownership (including the registration number, trademark owner, and location) on over 10 million trademarks from 1870 to the present. From these, OCE retained only entries where the nationality variable was listed as "US" (for American companies) or was missing, and where the registration date falls in the 1990 - 2016 range. Then, we standardized the owner name, city, and address variables using the *stnd_compname* and *stnd_address* packages described above. Similarly, we cleaned and standardized the alternative owner name variable, and removed prefixes including "doing business as" and "also known as". In cases where the standardized alternative name variable differed from the original owner name variable, OCE constructed additional observations containing the trademark data and the alternative owner name. Finally, we restricted the sample to include only owners at registration[6] when generating the final dataset of trademark-owner observations.[7]

3. The NETS database

I. Overview

The National Establishment Time Series (NETS) database by Walls & Associates uses yearly data sourced from Dun & Bradstreet (D&B) to provide detailed information on over 64 million

[3] For details on the methodology and use of these packages, see https://journals.sagepub.com/doi/abs/10.1177/1536867X1501500304.
[4] The final patent dataset is available on the COEAT-1 server as *all_pv_patents.dta* in the folder */data/project/pv_to_nets_2017/pv_preparation/*.
[5] The USPTO Trademark Case File Database is available at: https://www.uspto.gov/ip-policy/economic-research/research-datasets/trademark-case-files-dataset.
[6] Owners at registration are coded as ownership type 30 in the Trademark Case File Database.
[7] The final trademark dataset is available on the COEAT-1 server as *all_trademarks.dta* in the folder */data/project/tm_to_nets_2017/tm_preparation/*.

establishments for the years 1990-2016.[8] Among other items, the NETS database includes datasets with information on establishment locations, headquarters linkages, corporate structure, years of business activity, industry classification, employment, sales, and D&B credit ratings.

Each observation in the original D&B data source is an establishment–year pair, with establishments tracked over time using the "DUNS number" identifier (the longitudinal linkage is created by Walls & Associates). To match to patents and trademarks, we converted the NETS database to a panel-like structure (consisting of year ranges rather than years) for each name-address-DUNS number combination. This process included a number of steps: First, we linked the main NETS establishment-level dataset, *NETS2017_Company.dta*, to establishment-level name change records to obtain the time interval during which each establishment held any given name. Second, we adjusted address information and year ranges to reflect the establishment location changes on record. If we observed a name change and an address change in NETS for a particular DUNS number within a given year, we created a record for each name/address combination in that year. Third, we reshaped the data to incorporate trade name information for each observation in addition to the official name. Finally, we again applied the user-developed Stata packages *stnd_compname* and *stnd_address* to standardize the name, address, and city fields of all observations in preparation for matching to patents and trademarks. We provide details on each of these steps in the following sections.

II. Name changes

As described above, OCE processed the observations in the NETS database to account for name changes over time for each DUNS number. The first step of this procedure merged the main NETS dataset, *NETS2017_Company.dta* (which contains unique-DUNS-number-level data in each row for the most recently observed year in the NETS database, i.e. 2017), with the *NameSearch2017.dta* dataset via a one-to-many merge by DUNS number. The latter dataset contains all name changes linked to each DUNS number between 1990 and 2019[9]. For example, DUNS number 608893616 first appeared in the NETS data in 1991 as "LCOMP INC." In 1996, the company changed its name to the "CARLTON-BATES COMPANY," and kept this name until it ceased operations in 1998. In this instance, there are only two name records in the *NameSearch2017.dta* dataset (LCOMP INC and CARLTON-BATES COMPANY). The *NameSearch2017.dta* dataset contains roughly three million unique DUNS numbers that do not appear in the *NETS2017_Company.dta* dataset,

[8] Walls & Associates obtains their annual D&B source data in January each year reflecting the prior year's information. This means the 2017 release of the NETS database compiles all establishments through the end of 2016, precluding OCE from matching patents or trademarks filed past this point.
[9] While the NETS 2017 database only tracks new establishments through 2017, it tracks included establishments through 2019 in terms of potential name and address changes.

presumably due to the longer time period covered by *NameSearch2017.dta*. We retained only observations that matched both datasets.

The second step involved taking the output of the first step (above) and performing a many-to-one merge by DUNS number against the *NETS2017_Misc.dta* file, which contains, among other variables, each DUNS number's observed years of operations in the NETS database via the variables *firstyear* and *lastyear*. To create a modified panel of the name-year range combinations associated with a particular DUNS number, we used the variable *nameyear* from the *NameSearch2017.dta* dataset. This variable is "backward-facing," i.e. it indicates the final year a name was in effect for a given company. Returning to the example of DUNS number 608893616, we observe this establishment in the NETS database for the period of 1991 to 1998, with one observed name change in 1996. Specifically, the *nameyear* variable is populated as follows: one entry in 1996 - linked to the name of the establishment at first observation (i.e. "LCOMP INC"), and one entry in 1998 – linked to the name after the 1996 name change (i.e. "CARLTON-BATES COMPANY"). Thus, the establishment tracked by DUNS number 608893616 has two distinct names during its existence in the NETS database: "LCOMP INC" for 1991-1996 and "CARLTON-BATES COMPANY" for 1996-1998. We included both of these names as potential matches for patents and trademarks granted in 1996, but only the contemporaneous name for years before or after the name change.

OCE staff used these data to perform additional adjustments that ensure clear labelling of the year ranges and guarantee all data remained within the 1990-2016 range. While the 2017 version of the NETS database only tracks establishments that were in existence through the end of 2016, it tracks name changes for existing establishments beyond this point. In particular, for establishments with name change records where *nameyear* is greater than the latest year of NETS establishment coverage (i.e. 2016), the *nameyear* value was replaced with 2017. Therefore, an establishment with post-2016 name changes may be associated with multiple firm names listed under 2017.[10] We retained these records to allow for better match quality in cases of a lag in the official name change records. As a final step in generating name-year observations, we replaced the *lastyear* variable (last observed year for each DUNS number) with *nameyear*, and the *firstyear* variable (first observed year for each DUNS number) with the *nameyear* value of the previous name observation, sorted in chronological order. The resulting dataset contains the first and last year that a given name was associated with a particular DUNS number for each name-DUNS number combination.

[10] Our approach allowed patents and trademarks granted in the final year of NETS establishment coverage (2016) to match any post-2016 names used by a given establishment.

III. Address change and name change combinations

In addition to changing their names, many establishments also change their geographic location; records of these changes are provided in NETS file *AddressSearch2017.dta*. We used this information to create a modified panel that incorporates both name and address changes across time for any given DUNS number.

The first step in creating this modified panel was a many-to-one merge of the *AddressSearch2017.dta* dataset to the *NETS2017_Misc.dta* dataset by DUNS number. This allows one to link the first and last observed years and each known address for each establishment.

The file *AddressSearch2017.dta* has the same structure as the file *NameSearch.dta* discussed in subsection 3.II above, but instead focuses on address changes rather than name changes over time using the *yearaddress* variable. Similar to the process for incorporating name, we linked the main NETS dataset, *NETS2017_Company.dta*, to *AddressSearch2017.dta* via a one-to-many merge by DUNS number. As with name changes, address changes are tracked beyond 2016 for existing establishments: *AddressSearch2017.dta* contains address changes linked to each DUNS number between 1990 and 2019.[11]

For the second step in creating a panel of both name and address changes, we linked the file resulting from the previous step to *NETS2017_Misc.dta* via a many-to-one merge by DUNS number. We then performed two additional adjustments to ensure that we labeled all data clearly and that they remained within the 1990-2016 range. First, if an establishment has address-change records beyond the latest year of NETS establishment coverage (2016), we replaced the *yearaddress* value with 2017. As a consequence of this, establishments with post-2016 address changes may be associated with multiple firm addresses listed under 2017. Second, if DUNS numbers matched, we replaced the *lastyear* variable (last observed year for each DUNS number) with the *yearaddress* variable, and the *firstyear* variable (first observed year for each DUNS number) with the *yearaddress* value of the previous address observation, sorted in chronological order. The resulting dataset contained the first and last year that a given name was associated with a particular DUNS number for each address-DUNS number combination.

Next, we linked the address dataset constructed above (subsection 3.III) to the name change dataset (subsection 3.II) via a many-to-many join based on DUNS number. As a final preparatory step, in order to guarantee validation of all name and address combinations, we dropped all

[11] The *NameSearch2017.dta* file contains roughly 1.5 million address observations that do not appear in the *NETS2017_Company.dta* file, a discrepancy presumably due to the longer period covered by the *NameSearch2017.dta* data. However, for the merge, OCE limited observations considered to the intersection of the two datasets. As before, we retained only DUNS numbers that appeared in both datasets.

observations for which the first year of the observed address was larger than the last year of the observed name, or for which the first year of the observed address was smaller than the first year of the observed name. The combined dataset gives information for each establishment while accounting for the chronology of both name and address changes.

IV. Additional data preparation

The NETS database also provides trade names for a large portion of records. To incorporate these trade names, when provided, and list them alongside the official names, OCE staff reshaped the establishment name data so that establishments with a nonempty trade name have two records with different names (one official name and one trade name) but with otherwise identical content.

The file *NETS2017_AddressFirst.dta* contains first observed address record for each DUNS number, but this is sometimes inconsistent with the first observed address information from the *NETS2017_Move.dta* dataset. To ensure that the matching process remains robust to these discrepancies, we matched the data described in subsection 3.III, "Address Change and Name Change Combinations," to the *NETS2017_AddressFirst.dta* file, and retained both address observations in cases where first addresses differed.[12]

Finally, OCE once again employed the user-developed packages *stnd_compname* and *stnd_address* to parse and standardize the text of the establishment name, address, and city. The final NETS dataset contains 19 variables with over 104 million records identifying, among other things, the establishments, their names, and their location information across time.[13]

4. The Doherr SearchEngine

The Doherr SearchEngine (DSE) is software developed by Dr. Thorsten Doherr to implement heuristic matching in large databases based on fuzzy matching criteria, with specific optimizations for names and addresses. It prepares and uses two user inputted data tables, a "base table" and a "search table."[14] It starts by creating a base table registry that contains occurrence counts of each word in designated fields (called "search fields") of the base-table after performing some standardization steps such as transforming the fields to uppercase and removing punctuation. Next a user-supplied search table with similar search fields (e.g., name, city, state) is compared to the base table search fields, and the DSE software calculates the identification potential of each word using the inverse of its occurrence rate as its main matching criterion.[15] Different weights

[12] As a consequence, for the period between the first observed year and move/last observed year, there may be a small number of instances with two distinct address entries per observed name-time period combination.
[13] The file is available on the COEAT-1 server as *nets_panel_2017.dta* in the folder */data/NETS/NETS_formatted/*.
[14] Documentation for the Doherr SearchEngine is available at: https://github.com/ThorstenDoherr/searchengine.
[15] This ensures that commonly occurring words have a low identification potential.

can then be applied to the search fields (i.e., assignee name can be given a greater weight than the assignee city or state). Ultimately, the DSE generates an identification score for each match between the search table and the base table that is a weighted sum of the identification potential scores for each word in the fields utilized for the match. These scores identify the set of high-potential matches for each patent or trademark owner in the IP datasets.

5. Matching Procedure for Patents

The OCE staff matched patents granted to US-based organizations during the 1990 - 2016 period with the NETS database in five steps:

i) Performed an exact match between the PatentsView (PV) observations and NETS observations using Stata (see Section 3 for specifics on the NETS data applied here);
ii) Used the DSE to identify high-probability fuzzy matches between the residual patent/assignee pairs from step (i) and the NETS database;
iii) Manually matched large patent holders that had neither an exact nor a fuzzy (DSE) match;
iv) Associated each DUNS match to a headquarters-level HQ DUNS number, in order to report all final matches at the firm level rather than the establishment level; and
v) Assigned each patent/assignee pair to one firm-level match.

In the following subsections, we describe the above steps in detail.

I. PatentsView to NETS Exact Matches

Details of the patent data that were matched to the NETS database are in Section 2.I above. To account for records containing both an official company name and a tradename, OCE split the NETS dataset in two. The resulting datasets are *nets_formatted2017_company_name.dta* and *nets_formatted2017_trade_name.dta*,[16] named according to whether the tradename indicator was "0" (indicating that the name for a record corresponds to a NETS company name) or "1" (indicating that a name corresponds to a tradename), respectively. We required the split datasets because matching on tradenames alone frequently resulted in patents matched to many DUNS numbers. Since this complicates efforts to assign patent ownership to only one firm, OCE staff merged the patent data to the company name dataset first, followed by a merge to the tradename dataset for those patents still unmatched.

Using these datasets, we identified exact matches according to the following four steps:

i) We matched *all_pv_patents.dta* to *nets_formatted2017_company_name.dta* by *stn_name* (standardized name), *state*, and *city*;

[16] These files are both found on the COEAT-1 server in the folder, */data/project/pv_to_nets_2017/nets_formatted/*.

ii) For patents without a match in the previous step, we matched to *nets_formatted2017_company_name* by compressed *stn_name*, *state*, and compressed *city* (all white space taken out of *stn_name* and *city*);

iii) For patents without a match in the previous two steps, we matched to *nets_formatted2017_trade_name.dta* by *stn_name*, *state*, and *city*

iv) Finally, for any remaining patents, we matched to *nets_formatted2017_trade_name.dta* by compressed *stn_name*, *state*, and compressed *city*.

Output from these four steps was appended together to create the exact match file (for all years); namely, *exact_match_pv_nets_updated20200131.dta.*[17]

OCE then grouped these four types of exact matches described above by *grant_year* for each year in the range 1990-2016, with observations kept if the *grant_year* was greater than or equal to the *firstyear* variable and also less than or equal to the *lastyear* variable.[18] We retained only those matches for which the candidate firm match in NETS is considered active during the patent's grant year.[19]

II. PatentsView to NETS Fuzzy Matches

Following the exact matching steps, OCE generated lists of candidate matches for the remaining patents using the fuzzy matching capabilities of the Doherr SearchEngine (DSE). For each grant year, we created the base table registry by splitting the *nets_formatted.dta* dataset into yearly files, where each file contains only the active establishments (as determined by the *firstyear* and *lastyear* variables) within NETS.[20]

After analyzing the performance of the DSE fuzzy matching algorithm under different setting configurations, we found an iterative search over three stages to give the widest coverage and most reliable results. Each stage involved generating lists of candidate matches by using the DSE to link patents unmatched in any prior stage to establishments in the NETS database. To ensure that as many reasonable candidate matches as possible were identified, we used more permissive criteria in later stages of the process, applying these to patent-assignee observations that failed

---

[17] This file is located on the COEAT-1 server within the directory */data/project/pv_to_nets_2017/intermediate_files/*.
[18] Recall that the *firstyear* and *lastyear* generally correspond to the first year and last year that a particular DUNS number/address/company name combination is in the nets*_formatted2017.dta* dataset. However, for establishments that existed prior to 1990, *firstyear* is set to1989; similarly, for establishments that continue to operate past 2016, *lastyear* is set to 2017.
[19] Annual files containing exact PV-to-NETS matches are located on the COEAT-1 server in the directory */data/project/pv_to_nets_2017/intermediate_files/*, with matches for year YYYY stored in *exact_match_pv_nets_YYYY.dta*.
[20] The resulting matches are saved as *nets_formatted2017.dta* and are located on the COEAT-1 server at */data/project/pv_to_nets_2017/nets_formatted/*.

to generate matches in prior stages. This allowed us to identify candidate links for patents that would otherwise be unmatched without needing to lower the standard necessary for other patent-assignee observations. Table 1 below summarizes the processes and parameters OCE used for each of the different fuzzy-match stages.

Table 1. Settings for DSE Fuzzy Matching of PatentsView Data to NETS

DSE Fuzzy Matching Parameters	**Stage**		
	I	**II**	**III**
Hit Rate	95%	95%	95%
Name Weight	85%	90%	85%
City Weight	5%	0%	5%
State Weight	10%	10%	10%
Ngram-4: Name	No	No	Yes
Ngram-4: City	No	No	Yes
Darwinistic	Yes	Yes	Yes
Relative Search	Yes	Yes	Yes
Only Append New?	N/A	Yes	Yes
Feedback	10%	10%	No
Feedback Activation	2	2	N/A
Log Smoothing	No	No	Yes

Note: The Hit Rate is the weighted sum of similarity scores for the name, city, and state fields, using the weights specified in the Name Weight, City Weight, and State Weight entries, respectively. "Ngram_4" settings indicate that the fuzzy match is allowed to match based on four-letter sequences rather than the complete text of either name or city. The "Darwinistic" search option means that only candidate matches with the highest identification score, or tied for the highest identification score, are retained for consideration. The "Relative Search" option reallocates the weight for any missing search term to the remaining non-missing search terms. The "Only Append New" feature means that existing candidate matches from earlier stages are not repeated in later stages. The Feedback feature penalizes surplus words in the search table text records compared to the base table search terms. The reduction in the "hit" percentage score reflects the extent of the penalty. Logarithmic (log) smoothing indicates smoothing by the logarithmic inverse word frequency ratio, causing the DSE to be more conservative in identifying candidate matches.

To qualify as a candidate match, a pair of records needed to achieve a "hit rate" (similarity) of 95% in any stage. The DSE calculates the hit rate or similarity score as a weighted sum of similarities between one or more user-specified fields. We assigned the name field a weight of 85% of the total score, the state field 10%, and the city field 5%. To minimize the risk of false positive matches, OCE maintained this threshold throughout, and applied log smoothing to the final ngram-4 stage. Finally, as explained in the caption to Table 1, we applied feedback to penalize matches with surplus words when the number fuzzy match candidates exceeded the feedback activation level.

III. PatentsView to NETS Manual Matches

After identifying exact and fuzzy matches in the previous two steps, OCE staff isolated large patent holders from the remaining unmatched observations for manual assignment and review. While the vast majority of large patent holders did have at least one exact or fuzzy match, a small proportion remained unmatched after the above steps due to discrepancies in either names or addresses between the PatentsView and NETS databases. For this analysis, an organization is a large patent holder if an assignee-city-state combination received 50 or more patents in a given year. Those assignees were then manually matched to entries in the NETS database.[21]

IV. Disambiguating Exact and Fuzzy Matches

Prior to appending the exact, fuzzy, and manual matches together, OCE staff linked each patent belonging to one of the first two categories (exact and fuzzy matches) with a single, firm-level NETS entry. This process entailed two steps. First, we used information in NETS to link patent assignee information at the establishment level to firm-level headquarters information. Second, for cases where the list of candidate matches mapped to multiple distinct headquarters, additional information from NETS was used to determine a single best-fit headquarters for assignment. We describe these steps in turn below.

In order to convert the establishment-level assignee information to headquarters-level firm information, OCE used the NETS headquarters linkage information stored in the *NETS2017_HQs.dta* dataset, which contains yearly information on the parent establishment, or headquarters (HQ) of each NETS establishment from 1990-2016. Although most establishments have just one HQ, some firms have more complicated corporate structures with multiple levels of HQ listings. OCE staff linked each unique establishment-level entity from the matches in the previous steps to their highest-reported, corresponding yearly HQ using the variable, *hqdunsYY*[22]. As a result, each matched patent-assignee observation is linked to a firm-level identifier (*hqduns*) through its matched establishment-level DUNS number.

When the list of candidate matches for a patent contained multiple distinct headquarters or standalone establishments, OCE used additional information from the NETS database to identify a single most likely match. Variables used for this purpose included *estcat*, from the file *NETS2017_Misc.dta*, and *emp14c*, from *NETS2017_Emp.dta*. The variable *estcat* identifies whether a particular DUNS number is a headquarters or a standalone establishment. We use the *estcat* variable to classify the remaining matches into five broad categories, labeled A through E.

---

[21] The manual match results of patents from year YYYY are stored as *manual_match_YYYY.dta* within the directory */data/project/pv_to_nets_2017/intermediate_files/*.

[22] The *YY* in *hqdunsYY* refers to the relevant two-digit year recorded in the data.

Category A identifies patent-assignee pairs matched to a unique HQ DUNS number across all Doherr SearchEngine results, such that no further disambiguation is necessary. Category B corresponds to patent/assignee pairs with a single HQ DUNS number and one or more stand-alone DUNS numbers. Category C is the set of patent-assignee pairs for which all candidate matches consist of only stand-alone DUNS numbers. Category D consists of patent-assignee pairs with two or more HQ DUNS numbers as well as possible stand-alone DUNS numbers. Category E contains patent-assignee pairs for which the HQ DUNS number did not exist in the *NETS2017_Company.dta* file. Across these categories, the variable *match_flag* provides more granular detail regarding the exact method used to disambiguate the patent-assignee candidate matches. Table 2 provides definitions for the different levels of the *match_flag* variable. Within the context of entity resolution, the above process is a second level of clustering that OCE generates on top of the establishment-level disambiguation used for the initial match between PatentsView and NETS.[23]

[23] OCE refers to this process informally as "rolling up." Occasionally, the *lastyear* (i.e., the last year that a given DUNS number was active) variable in the *NETS_formatted.dta* dataset contradicted the information from *NETS2017_Misc.dta*. This is because OCE staff constructed the dataset with the intent to include every instance of name/address change. However, this introduced discrepancies between the constructed dataset and other NETS files, including *NETS2017_Misc.dta*. In order to preserve as much information as possible, OCE staff created two groups. In the first group, the rows corresponding to a patent/assignee pair were retained if the *lastyear* variable in the *NETS_formatted.dta* dataset was less than or equal to the *lastyear* variable in the *NETS2017_Misc.dta* file. These patent-assignee pair observations were then mapped to the appropriate *hqdunsYY* variable using the *NETS2017_HQs.dta* dataset. In the second group, the *lastyear* variable from *NETS_formatted* was greater than the *lastyear* variable in the *Misc* dataset for all rows; all of these matches were retained. These were then linked to each year's HQ DUNS number using *NETS2017_HQCompany* instead, which contained the last observed record.

Table 2: Definitions for the match_flag variable.

Flag	# of Obs.	Description
		Category A: Patent-assignee pairs attached to a single DUNS number
1	1,692,997	Patent-assignee pair assigned to unique HQ DUNS number in the previous step – no further action necessary
		Category B: Patent-assignee pairs match to a single headquarters, as determined by the *estcat* variable (others are standalones)
2	244,743	Assigned to a DUNS number identified as the headquarters
		Category C: Patent-assignee pairs match only to standalone DUNS numbers, as determined by *estcat* variable
3	1,636	Isolate the patent-assignee pairs from bucket 2 **with** the same address and assign them to the DUNS number with real employment information according to the *Empc* variable in NETS.
4	8,527	Isolate the patent-assignee pairs from bucket 2 **without** the same address and assign them to the DUNS number with non-missing and non-zero employment information according to the *Empc*[24] variable in NETS.
5	12,226	For the remaining cases from Category C, assign them randomly to one of the DUNS.
		Category D: Patent-assignee pairs match to multiple headquarters, as determined by *estcat*. (OCE then dropped non-HQ DUNS numbers)
6	49,718	Identify the patent-assignees from Category D with 25 or more patents and assign them manually.
7	1,465	Isolate the patent-assignee pairs from Category D **with** the same address and assign them to the DUNS number with real employment information according to the *Empc* variable in NETS.
8	4,490	Isolate the patent-assignee pairs from Category D **without** the same address and assign them to the DUNS number with real employment information according to the *Empc* variable in NETS.
9	9,504	For the remaining cases from Category D, assign them randomly to one of the DUNS numbers.
		Category E: If the HQ DUNS number does not exist in the *NETS2017_Company* file, use the original DUNS number instead
10	13,613	Identify the patent-assignee pairs from Category E matched to only one DUNS number and assign them to this.

[24] The *NETS2017_Emp* dataset has variables called *empcYY* (where *YY* corresponds to a two-digit year) that indicate whether the employment number at a particular establishment for a given year is the exact figure, an estimate provided by Dun & Bradstreet, or an estimate provided by Walls & Associates.

6. Matching Results for Patents

I. Final Dataset

The final dataset matching patents to NETS contains a total of 2,038,919 observations over the years 1990-2016.[25] Table 3 lists the variables in the dataset. Of these, *pv_pat_no* is the PatentsView patent number, and *dunsnumber* is the DUNS number of the ultimate organizational parent (headquarters or HQ). The standardized name from PatentsView is *stn_name*, while company and tradename are the NETS official name and the NETS tradename, respectively. Observations that we manually matched to their ultimate parent are indicated by the variable *manual_match*, while fuzzy matches via the SearchEngine matches are indicated by *SE_Match*. The establishment category (HQ, branch, or standalone) is reported by *estcat*. Finally, *match_flag* provides more information on the exact matching method, as described in Table 2 above.

Table 3: Key variables and definitions for PatentsView to NETS match.

Variable Name	Description
city	City (NETS)
company	NETS official name
dunsnumber	Ultimate Parent (HQ) DUNS number
estcat	Last Type of Location
grant_year	Grant Year
manual_match	Indicator for Manual Matches
match_flag	Match Type Code (See Table)
pat_no	Patent Number
pv_city	City (PatentsView)
pv_state	State (PatentsView)
SE_match	Indicator for SearchEngine Match
state	State (NETS)
stn_name	Standardized Assignee Name
tradename	NETS trade name

II. Matching Statistics for Patents

Table 4 provides basic summary statistics for the PatentsView to NETS matching results for each of the years in 1990 – 2016. The second column reports the total number of patents considered in each year, while the third column gives the number of those patents that we successfully linked

[25] The final patent dataset, *final_pv_nets_match.dta*, is in the directory */data/project/pv_to_nets_2017/final_files/* on the COEAT-1 server.

to an entity in the NETS database. The final column is the ratio of the third column and the second column. The minimum matching rate occurs in 1999 with 84.9%. All other years have a match rate of at least 87%; the aggregate average match rate is 88.5%.

Table 4: Number of patents successfully matched by year.

Year	Patents	Matched	Fraction
1990	40,423	35,910	88.8%
1991	44,292	38,599	87.1%
1992	45,431	40,489	89.1%
1993	47,525	42,017	88.4%
1994	50,496	44,617	88.4%
1995	50,393	44,715	88.7%
1996	55,002	48,334	87.9%
1997	56,864	50,511	88.8%
1998	73,869	65,335	88.4%
1999	77,290	65,603	84.9%
2000	80,282	71,311	88.8%
2001	83,775	74,391	88.8%
2002	83,771	73,459	87.7%
2003	85,253	75,210	88.2%
2004	82,309	73,360	89.1%
2005	72,601	65,251	89.9%
2006	90,074	81,239	90.2%
2007	82,793	74,616	90.1%
2008	82,465	73,926	89.6%
2009	85,739	76,380	89.1%
2010	109,375	98,925	90.4%
2011	111,064	98,121	88.3%
2012	123,890	108,527	87.6%
2013	137,946	121,246	87.9%
2014	149,939	130,780	87.2%
2015	147,616	131,063	88.8%
2016	152,953	134,984	88.3%
Total	**2,303,430**	**2,038,919**	**88.5%**

As an initial quality check, OCE staff manually examined the match for patent assignees with 100 or more patents in each grant year in order to determine if they were matched to the correct firm. We manually corrected inaccurate matches.[26] Then, as a further quality check, OCE selected and examined the assignee-to-firm match for at least 200 assignees for the years 2000-2016, and 50 assignees each for the years in 1990-1999. This allowed for computation of the false positive rate

[26] The code contains the full list of firms for which OCE staff performed manual matches to the NETS data.

for the matches both at the assignee level, as well as at the patent-weighted level (taking into account the number of patents owned by the assignees examined in the first step). Table 5 summarizes the findings of this quality check by at the assignee level and highlights a true-positive match rate of greater than 90% within the manually-examined sample.

Table 5: Assignee-level false positive rate for PatentsView to NETS match

Match	N	Percent
False Positive	200	9.52
True Positive	1,900	90.48
Total	2,100	100.00

Table 6 shows the false positive rate at the patent-weighted level broken down by whether the initial match was exact or fuzzy (i.e. from the DSE). The table also shows the rates of false and true positives in italicized text for each matching process. Although the false positive rate for the fuzzy matches is slightly higher than for the exact matches, the bulk of patents in the manual-examination sample come from exactly-matched assignees (95.8%), leading to an overall true-positive rate of approximately 93%.[27]

Table 6: Patent-level false positive rate for PatentsView to NETS match

	Exact Match	Fuzzy Match	Total
False Positive	1,003	1,402	2,405
	4.21%	*14.30%*	*7.16%*
True Positive	22,808	8,399	31,207
	95.79%	*85.70%*	*92.84%*
Total	23,811	9,801	33,612

7. Matching Procedure for Trademarks

I. Overview

This section describes the process used to match trademarks in the Trademark Case File (TMCF) database to headquarters-level entries in the NETS database. Although there are many similarities to the procedure followed for matching patents to the NETS database, the experience gained by OCE staff in the PatentsView to NETS matching process and differences in the characteristics of the TMCF and PatentsView databases led to some noteworthy changes in the Trademarks to NETS matching process. We describe these changes below.

[27] Note that the patent-level true positive rate of 93% is slightly higher than the assignee-level rate of 90.5%. This suggests that our matching methodology is more effective for assignees with larger numbers of patents.

II. Trademarks to NETS Exact and Fuzzy Matches

In contrast to the procedure used for matching patents to the NETS database, the Trademarks to NETS matching process did not include a preliminary round of filtering exact matches. Instead, OCE began by using the Doherr SearchEngine (DSE) to construct lists of candidate matches for all trademarks. The DSE settings and parameters used were identical to those followed for constructing the PatentsView to NETS candidate match lists (see Table 1). Notably, since the Darwinistic setting means that the DSE matching process retained only candidate matches with (or tied for) the highest score, the DSE results will return the exact match and only the exact match in cases where one exists. This modification obviated the need to handle instances of exact matches as a separate step. In order to manage the number of candidate matches for each trademark owner, we generated annual candidate lists for each trademark registration year from 1990 through 2016. For each year's candidates, we used the full set of name-city-state observations in NETS that were active within one year of the desired trademark registration year.

III. Disambiguating Trademarks to NETS Matches

After identifying candidate matches, OCE followed a procedure similar to that used in PatentsView to NETS above to associate each trademark with a single most-likely match. For this disambiguation, we again distinguish between HQ-DUNS numbers, which refer to the top-level headquarters associated with any given DUNS number, and Establishment-DUNS numbers, which refer to the potential NETS matches returned by the DSE.[28] OCE used the list of questions below to disambiguate multiple trademark candidate matches.

1. How many potential matches did the Search Engine return?
   a) If exactly 1: assign the HQ-DUNS number as the match
   b) If exactly 0: categorize the TM case file as unmatched
   c) If more than 1: go to step (2)
2. How many HQ-DUNS numbers are associated with the set of potential matches from the Search Engine during the focal year?
   a) If exactly 1: assign the HQ DUNS number as the match
   b) If exactly 0: go to step (3) while keeping all potential matches from this step.
   c) If more than 1: go to step (3) using only the HQ-DUNS numbers as potential matches
3. How many potential matches have positive employment in the focal year? [Note: use the employment of the HQ-DUNS]

[28] As with PatentsView to NETS, an Establishment-DUNS number may identify a standalone establishment, a branch of a multi-establishment entity, or the headquarters (HQ) of a multi-establishment entity. For both standalone establishments and for headquarters, the HQ-DUNS number is identical to the establishment-DUNS number.

   a) If exactly 1: assign the HQ-DUNS number as the match
   b) If exactly 0: go to step (4) while keeping all potential matches from this step
   c) If more than 1: go to step (4) using only the positive-employment observations as potential matches
4. How many potential matches are in the same zip code as the TM case file? [Note: use the establishment-DUNS address, not that of the HQ-DUNS]
   a) If exactly 1: assign the HQ-DUNS number as the match
   b) If more than one: keep only these observations and proceed to step (7)
   c) If zero: go to step (5)
5. How many potential matches are within 10 miles of the zip code of the TM case file? [Note: use the establishment-DUNS address, not that of the HQ-DUNS]
   a) If exactly 1: assign the HQ-DUNS number as the match
   b) If more than one: keep only these observations and proceed to step (7)
   c) If zero: go to step (6)
6. How many potential matches are within 25 miles of zip code of the TM case file? [Note: use the establishment-DUNS address, not that of the HQ-DUNS]
   a) If exactly 1: assign the HQ-DUNS number as the match
   b) If more than one: keep only these observations and proceed to step (7)
   c) If zero: go to step (7) with all potential matches
7. How many potential matches have at least half as much employment as the highest-employment potential match? [Note: use the employment for the HQ-DUNS]
   a) If exactly 1 (i.e. the highest-employment match has more than twice as many employees as the next-highest alternative): assign the HQ-DUNS number as the match
   b) If more than one: keep only these observations and proceed to step (8)
8. Choose the modal HQ-DUNS of all remaining potential matches
   a) In the event of a tie, assign the HQ-DUNS number with the earliest *yearstart*, i.e. founding year among the set of modal duns matches
   b) In the event of a further tie, assign to the highest-employment HQ-DUNS number among the remaining set
   c) In the event that both founding year and employment are tied, assign the highest HQ-DUNS number as the match.

To facilitate replication, the final tiebreaker is choosing the larger DUNS number. The largest DUNS number is preferred over the smallest because higher DUNS numbers correlate (weakly) with older establishments and those with higher employment.

8. Matching Results for Trademarks

I. Final Dataset

The final Trademarks to NETS dataset contains a total of 2,626,603 observations over the years 1990 – 2016.[29] Table 7 lists the variables in the dataset. *TM_ID* is the trademark registration number. Similarly to the PatentsView to NETS matched dataset, the variable *dunsnumber* is the DUNS number of the ultimate organizational parent, and the establishment category of this headquarters is reported by the *estcat* variable. The DUNS number of the establishment matched to the TMCF entry by the DSE is reported separately as *orig_SE_dunsnumber*. As in the PatentsView to NETS matched dataset, the standardized company name from the TMCF is reported as *stn_name*, while *company* and *tradename* are the NETS official name and the NETS tradename, respectively. The type of match between the TMCF and headquarters is given by the variable *match_code*, which Table 8 describes in detail.

Table 7: Key variables and definitions for Trademarks to NETS match

Variable Name	Description
city	City (NETS)
company	NETS official name
dunsnumber	Ultimate Parent (HQ) DUNS number
estcat	Last Type of Location
grant_year	Grant Year
match_code	Match Type Code (See Table)
orig_SE_dunsnumber	Establishment DUNS number
state	State (NETS)
stn_name	Standardized Assignee Name
tm_city	City (TMCF)
TM_ID	Trademark registration number
tm_state	State (TMCF)
tradename	NETS trade name

[29] The final dataset is stored in the directory */data/project/tm_to_nets/2017/Output/* on the COEAT-1 server as *tm_nets_final_file.dta*.

Table 8: Match codes and definitions for Trademarks to NETS match

Match Code	# of Obs.	Description
10000	2,027,178	Unique SE Match
20000	317,686	Unique SE HQ
70000	73,169	SE Match to >1 HQ at Same Address, Employment Tiebreaks
80000	65,479	SE Match to >1 HQ at Different Addresses, Employment Tiebreaker
90500	21,568	Extended Tiebreakers: Unique Match in ZIP Code
90600	6,982	Extended Tiebreakers: Unique Match within 10 Miles
90700	5,227	Extended Tiebreakers: Unique Match within 25 Miles
90801	26,980	Extended Tiebreakers: > 1 in 25 Miles, by Employment
90802	34,884	Extended Tiebreakers: > 1 in 25 Miles, by Modal DSE Result
90803	13,985	Extended Tiebreakers: > 1 in 25 Miles, by Earliest Starting Year
90804	2,650	Extended Tiebreakers: > 1 in 25 Miles, by Highest Employment
90805	1,512	Extended Tiebreakers: > 1 in 25 Miles, by Highest DUNS
90811	10,023	Extended Tiebreakers: 0 in 25 Miles, by Employment
90812	5,886	Extended Tiebreakers: 0 in 25 Miles, by Modal DSE Result
90813	11,164	Extended Tiebreakers: 0 in 25 Miles, by Earliest Starting Year
90814	1,055	Extended Tiebreakers: 0 in 25 Miles, by Highest Employment
90815	1,175	Extended Tiebreakers: 0 in 25 Miles, by Highest DUNS

II. Matching Statistics for Trademarks

Table 9 presents trademark counts by year for each stage of the process. The column TMCF indicates the number of unique case numbers from the Trademark Case Files that OCE staff attempted to match to the NETS database using the DSE. The next column, DSE, counts the distinct case numbers for which the DSE was able to identify at least one candidate match and for which the organization name was not missing. The following column reports the number of trademarks for which the disambiguation process was able to select a match from the list of candidates. The last two columns of the table report the ratios of the TMCF and DSE columns and the DSE and Matched columns, respectively. With the exception of 1990, the first year of data, the DSE was able to identify a candidate match for at least 70% of observations in each year of the data. Although this statistic does represent a lower success rate in absolute terms than the corresponding figure for patents, it is impressive in light of the relatively higher density of small standalone establishments within the TMCF, since these small establishments are precisely those that are least likely to have complete coverage in the NETS database. The final column shows that, for each year, over 98% of trademarks treated by the DSE were successfully disambiguated.

Table 9: Yearly counts for Trademarks to NETS match.

Year	TMCF	DSE	Matched	DSE/TMCF	Matched/DSE[30]
1990	60,753	26,817	26,440	44.14%[31]	98.59%
1991	45,598	34,472	33,886	75.60%	98.30%
1992	82,297	62,353	61,273	75.77%	98.27%
1993	78,699	59,668	58,729	75.82%	98.43%
1994	63,201	47,982	47,231	75.92%	98.43%
1995	83,217	62,492	61,483	75.10%	98.39%
1996	92,563	69,628	68,913	75.22%	98.97%
1997	110,782	83,428	82,543	75.31%	98.94%
1998	99,766	74,660	73,767	74.84%	98.80%
1999	99,887	74,210	73,347	74.29%	98.84%
2000	123,840	91,393	90,390	73.80%	98.90%
2001	120,800	89,027	88,108	73.70%	98.97%
2002	163,614	121,269	120,333	74.12%	99.23%
2003	142,842	106,124	105,306	74.29%	99.23%
2004	123,341	91,835	91,165	74.46%	99.27%
2005	131,162	97,501	96,800	74.34%	99.28%
2006	162,806	120,957	120,249	74.30%	99.41%
2007	178,274	132,340	131,585	74.23%	99.43%
2008	199,282	147,873	146,996	74.20%	99.41%
2009	177,291	127,179	126,427	71.73%	99.41%
2010	162,674	122,131	121,526	75.08%	99.50%
2011	175,563	130,589	129,822	74.38%	99.41%
2012	180,714	132,101	131,374	73.10%	99.45%
2013	181,617	131,625	130,842	72.47%	99.41%
2014	187,661	135,933	134,998	72.44%	99.31%
2015	188,445	136,365	135,283	72.36%	99.21%
2016	190,966	138,825	137,787	72.70%	99.25%
Total	3,607,655	2,648,777	2,626,603	73.42%	99.16%

[30] The instances where it was not possible to disambiguate candidate match lists from the DSE were caused by all candidate matches being more than 50 miles from the address in the TMCF.

[31] The match rate in 1990 is significantly lower than other years, despite no difference in methodology. This is likely due to 1990 being the first year of the NETS database, with less comprehensive coverage of smaller establishments.

To estimate the accuracy of Trademarks to NETS matches, OCE staff implemented a two-part procedure. First, we assumed that matches with identical (standardized) entity name, city, and state across the Trademarks case file dataset and the NETS database were accurately identified. Second, we stratified the remaining matches by trademark grant year and then randomly sampled, resulting in 797 trademark case files selected for manual examination. As with the PatentsView to NETS match, this allows for computation of the false positive rate for our matches at the level of individual trademark case files. We found that for fuzzy matches, 548 / 797 (68.76%) were correctly identified and were therefore classified as true positives. Because only non-exact matches were included in the manual examination of the Trademarks to NETS matched dataset, the above percentages need to be adjusted to calculate the overall match rate.[32] Table 10 below presents overall rates and 95% confidence intervals for both false-positive and true-positive matches. Specifically, we estimate an overall true-positive match rate in excess of 86%, driven in large part by the fact that the majority of matched trademarks are the result of exact matching.

Table 10: Trademark-level false positive rate for Trademarks to NETS match

	Exact Match	Fuzzy Match	Total	Confidence Interval
False Positive	0%	31.24%	13.93%	(12.5%, 15.4%)
True Positive	100%	68.76%	86.07%	(84.6%, 87.5%)
Total	1,455,241 *55.40%*	1,171,362 *44.60%*	2,626,603 *100%*	

9. Conclusion

This paper documents the matching procedures OCE used to generate crosswalks between the USPTO's patent and trademark datasets and the 2017 National Establishment Time Series (NETS) database. The final patent dataset is able to match 88% of patents granted from 1990 through 2016, with an estimated true-positive match rate of 92%. Over the same time period, we matched 73% of trademarks to the NETS database, with an estimated true-positive match rate of 86%. The differences in match rates reflect the distinct natures of patents and trademarks: patents require an extended and costly application process, and are therefore usually pursued by larger and better-funded entities. By contrast, trademarks offer a relatively fast and affordable form of intellectual property, and are used by a much broader range of entities including small businesses with limited visibility. Because of this, the representative patent assignee is more likely to be included in the NETS database than the representative trademark owner. Even with these caveats, the crosswalks for both patent and trademarks offer a foundation for a wide range of future work tracking intellectual property within establishments over time.

[32] In additional testing, there was no significant trend in false positive rates over time. The calculations in Table 10 therefore assume a constant false-positive rate across the sample period of 1990-2016.

The most direct application of these data would be to track the intellectual property activity of firms, especially those that are publicly-traded and offer detailed financial statements each quarter. However, the more significant benefit of matching to the NETS database is that its coverage expands well beyond publicly-traded firms, including privately-held corporations, small businesses, new ventures, educational institutions, and non-profit and government organizations. This breadth of coverage provides a comprehensive view of the intellectual property ecosystem, and has the potential to address a wide range of research and policy questions related to the use of patents and trademarks in society.