

# UVA Darden Global Corporate Patent Dataset: Construction and Features

April 2019

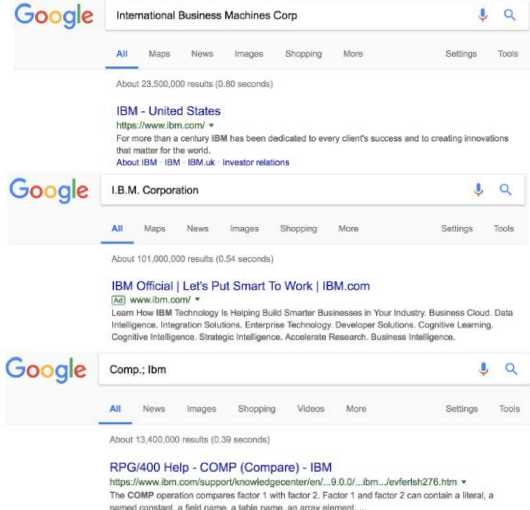
This document discusses the construction of a new patent-firm linked database: *Global Corporate Patent Dataset*. The dataset covers patents awarded by the U.S. Patent and Trademark Office (USPTO) to publicly listed firms internationally—those that are covered by the [S&P Compustat Global](#) database—in the 1980-2017 period.

Mapping of patent assignees to companies covered by external databases is a complex task. This is the case for two main reasons. First, companies' name strings that can be extracted from patent application and grant documents are not the companies' exact legal names and they are also not standardized. This means that there is no identifier that would uniquely flag each patent assignee in the available patent databases. In fact, for companies that have been active in innovation over a long period of time, there is typically a large number of different patent assignee strings that represent variants of the same company's name. Second, companies often file for patents through subsidiaries, where the subsidiaries' names typically do not correspond to their parent companies' names, and quite often the subsidiary and parent company names have little or no common component.

In prior work, researchers typically used fuzzy-string matching techniques to create links between patent assignee strings extracted from patent documents and companies' name strings extracted from external databases. Most work focused on matching USPTO granted patents to U.S. listed firms (covered by [S&P Compustat North America](#) or [CRSP](#)), for example, [NBER Patent Data Project](#) or [Kogan, Papanikolaou, Seru, and Stoffman \(2017\)](#). [Graham, Grim, Islam, Marco, and Miranda \(2018\)](#) match USPTO granted patents to administrative databases of firms and workers housed at the U.S. Census Bureau. They use inventor information in addition to the patent assignee firm name to improve on previous efforts linking patents to firms. [Bena, Ferreira, Matos, and Pires \(2017\)](#) use fuzzy-string matching techniques to match USPTO patents to firms internationally. [Bena, Ferreira, Matos, and Pires \(2017\)](#) highlight the importance of non-U.S. based companies for patenting and innovation activities more broadly: (i) Combined R&D spending of non-U.S. firms exceeded that of U.S. firms over 2000s. (ii) In recent years, USPTO granted more patents to non-U.S. firms than U.S. firms.

The truly global nature of innovation calls for a comprehensive data on patenting by international firms. To this end, the [Batten Institute](#) at the UVA Darden funded a project to create the *Global Corporate Patent Dataset*. The mapping of patent assignees to companies internationally is especially challenging because companies' names are in many languages and name strings contain suffices denoting different legal forms of incorporation according to local corporate laws. In addition, large non-U.S. companies are typically organized as conglomerates or pyramids with numerous member firms, subsidiaries, and affiliates. To overcome this challenge, when creating the *Global Corporate Patent Dataset*, we complement the fuzzy-string matching techniques used by [Bena, Ferreira, Matos, and Pires \(2017\)](#) (see the [Internet Appendix](#) for the detailed description of the fuzzy-string matching methodology in the international setting) with disambiguation of companies' names utilizing the capability of the internet

search engines. Specifically, we use the internet search engines to search for the patent assignee strings to obtain the domain name associated with each string. For example, [google.com](https://www.google.com) correctly returns [www.ibm.com](https://www.ibm.com) for patent assignee strings such as: “International Business Machines Corp”, “I.B.M. Corporation”, or “Comp.; Ibm”. We then correctly match patent assignee strings to companies in external databases directly using the companies’ domain names. This method is effective as it reliably works in any language and in most countries around the world.

Example 1: IBM																																	
Fuzzy string matching:	Novel approach:																																
<p>Firm name string in Compustat: conml = “International Business Machines Corp”</p> <table> <thead> <tr> <th>assg_name</th><th>dist</th></tr> </thead> <tbody> <tr><td>International Business Machines Corp</td><td>0.0000</td></tr> <tr><td>International Business Machiness Corporation</td><td>0.0625</td></tr> <tr><td>International Business Machines Corporation Corporation</td><td>0.0909</td></tr> <tr><td>International Business Machines Corporation</td><td>0.1250</td></tr> <tr><td>International Business Machines</td><td>0.1333</td></tr> <tr><td>International Business Machines Corporation, A Corporation</td><td>0.1667</td></tr> <tr><td>International Business Machines Corporation (IBM)</td><td>0.2105</td></tr> <tr><td>Ibm Corporation (International Business Machines)</td><td>0.2308</td></tr> <tr><td>International Business Machines Corporation Armonk Ny</td><td>0.2683</td></tr> <tr><td>International Business Machines Corporation New Orchard Road</td><td>0.3750</td></tr> <tr><td>International Business Machines Corporation, Armonk, New York 10504</td><td>0.4340</td></tr> <tr><td>I.B.M. Corporation</td><td>0.7895</td></tr> <tr><td>Ibm Corp.</td><td>0.8824</td></tr> <tr><td>Comp.; Ibm</td><td>0.9737</td></tr> <tr><td>Ibm</td><td>1.0000</td></tr> </tbody> </table>	assg_name	dist	International Business Machines Corp	0.0000	International Business Machiness Corporation	0.0625	International Business Machines Corporation Corporation	0.0909	International Business Machines Corporation	0.1250	International Business Machines	0.1333	International Business Machines Corporation, A Corporation	0.1667	International Business Machines Corporation (IBM)	0.2105	Ibm Corporation (International Business Machines)	0.2308	International Business Machines Corporation Armonk Ny	0.2683	International Business Machines Corporation New Orchard Road	0.3750	International Business Machines Corporation, Armonk, New York 10504	0.4340	I.B.M. Corporation	0.7895	Ibm Corp.	0.8824	Comp.; Ibm	0.9737	Ibm	1.0000	
assg_name	dist																																
International Business Machines Corp	0.0000																																
International Business Machiness Corporation	0.0625																																
International Business Machines Corporation Corporation	0.0909																																
International Business Machines Corporation	0.1250																																
International Business Machines	0.1333																																
International Business Machines Corporation, A Corporation	0.1667																																
International Business Machines Corporation (IBM)	0.2105																																
Ibm Corporation (International Business Machines)	0.2308																																
International Business Machines Corporation Armonk Ny	0.2683																																
International Business Machines Corporation New Orchard Road	0.3750																																
International Business Machines Corporation, Armonk, New York 10504	0.4340																																
I.B.M. Corporation	0.7895																																
Ibm Corp.	0.8824																																
Comp.; Ibm	0.9737																																
Ibm	1.0000																																

Most importantly, in many cases, this method correctly identifies member firms of conglomerates and corporate pyramids, which is crucial since most innovations are created by multinational companies often with operations in many countries and multiple R&D centres. In fact, business groups—multiple tiers of partially-owned listed affiliates and fully-owned private affiliates—is a dominant organizational form around the world outside of the U.S. ([Kandel, Kosenko, Morck, and Yafeh \(2018\)](#)). [Thoma, Torrisi, Gambardella, Guellec, Hall, and Harhoff \(2010\)](#) describe patenting by subsidiaries as one of the key challenges in linking patents to companies internationally. When applying this method, for example, [google.com](https://www.google.com) returns domain name [abbott.com](https://www.abbott.com) for patent assignee strings such as: “ABBOTT LABORATORIES”, “Abbott GmbH & Co.”, or “Abbott Healthcare Products, B.V.”, correctly identifying that these entities are subsidiaries of the same parent company. In most cases, this method therefore matches different patent assignee strings that belong to firms/affiliates/subsidiaries from the same corporate business structure to the parent company of this structure.

Compared to the fuzzy-string matching techniques, another advantage of this method is the ability to correctly disambiguate companies’ names that are quite similar. While strings “ABBOTT LABORATORIES” and “ATT LABORATORIES” appear to be very similar when compared using common string-distance metrics (for example, the Levenshtein distance), [google.com](https://www.google.com) returns correct domain name [abbott.com](https://www.abbott.com) for the former and [att.com](https://www.att.com) for the latter.

Example 2: Abbott Laboratories	
Fuzzy string matching:	Novel approach:

<p>Firm name string in Compustat: conml = "Abbott Laboratories"</p> <table> <tr> <th>assg_name_STD</th><th>dist_STD</th></tr> <tr><td>ABBOTT LABORATORIES</td><td>0.0000</td></tr> <tr><td>ABBOOTT LABORATORIES</td><td>0.0625</td></tr> <tr><td>ABBOT LABORATORIES</td><td>0.0667</td></tr> <tr><td>ABOTT LABORATORIES</td><td>0.0667</td></tr> <tr><td>ABBOTT LABATORIES</td><td>0.1250</td></tr> <tr><td>ABBOTT LABORAIOIES</td><td>0.1250</td></tr> <tr><td><b>ATT LABORATORIES</b></td><td><b>0.1333</b></td></tr> <tr><td>ABBOTT LABORATORIES LTD</td><td>0.1667</td></tr> <tr><td>ABBOTT LABORTORIES</td><td>0.1875</td></tr> <tr><td>ABBOTT LABORATIES</td><td>0.1875</td></tr> <tr><td>THE ABBOTT LABORATORIES</td><td>0.2105</td></tr> <tr><td>ABBOTT LABORATORIES INC</td><td>0.2105</td></tr> <tr><td>ABBOTT LABORATORIES INC</td><td>0.2105</td></tr> <tr><td>ABBOTT LABORAOTRIES</td><td>0.2353</td></tr> <tr><td>ABBOTT LABORATORY</td><td>0.2500</td></tr> <tr><td>ABBOTT LABORAIOIRES</td><td>0.2778</td></tr> <tr><td><b>BBJ LABORATORIES</b></td><td><b>0.2941</b></td></tr> <tr><td><b>ACT LABORATORIES</b></td><td><b>0.2941</b></td></tr> <tr><td><b>BABB LABORATORIES</b></td><td><b>0.2941</b></td></tr> </table>	assg_name_STD	dist_STD	ABBOTT LABORATORIES	0.0000	ABBOOTT LABORATORIES	0.0625	ABBOT LABORATORIES	0.0667	ABOTT LABORATORIES	0.0667	ABBOTT LABATORIES	0.1250	ABBOTT LABORAIOIES	0.1250	<b>ATT LABORATORIES</b>	<b>0.1333</b>	ABBOTT LABORATORIES LTD	0.1667	ABBOTT LABORTORIES	0.1875	ABBOTT LABORATIES	0.1875	THE ABBOTT LABORATORIES	0.2105	ABBOTT LABORATORIES INC	0.2105	ABBOTT LABORATORIES INC	0.2105	ABBOTT LABORAOTRIES	0.2353	ABBOTT LABORATORY	0.2500	ABBOTT LABORAIOIRES	0.2778	<b>BBJ LABORATORIES</b>	<b>0.2941</b>	<b>ACT LABORATORIES</b>	<b>0.2941</b>	<b>BABB LABORATORIES</b>	<b>0.2941</b>	
assg_name_STD	dist_STD																																								
ABBOTT LABORATORIES	0.0000																																								
ABBOOTT LABORATORIES	0.0625																																								
ABBOT LABORATORIES	0.0667																																								
ABOTT LABORATORIES	0.0667																																								
ABBOTT LABATORIES	0.1250																																								
ABBOTT LABORAIOIES	0.1250																																								
<b>ATT LABORATORIES</b>	<b>0.1333</b>																																								
ABBOTT LABORATORIES LTD	0.1667																																								
ABBOTT LABORTORIES	0.1875																																								
ABBOTT LABORATIES	0.1875																																								
THE ABBOTT LABORATORIES	0.2105																																								
ABBOTT LABORATORIES INC	0.2105																																								
ABBOTT LABORATORIES INC	0.2105																																								
ABBOTT LABORAOTRIES	0.2353																																								
ABBOTT LABORATORY	0.2500																																								
ABBOTT LABORAIOIRES	0.2778																																								
<b>BBJ LABORATORIES</b>	<b>0.2941</b>																																								
<b>ACT LABORATORIES</b>	<b>0.2941</b>																																								
<b>BABB LABORATORIES</b>	<b>0.2941</b>																																								
<p>Firm name string in Compustat: conml = "Abbott Laboratories"</p> <table> <tr> <th>assg_name</th><th>dist</th></tr> <tr><td>Abbott Laboratories</td><td>0.0000</td></tr> <tr><td>Abbott Biotherapeutics Corporation</td><td><b>0.7429</b></td></tr> <tr><td>Abbott Gmbh &amp; Co.Kg A Corporation</td><td><b>0.7500</b></td></tr> <tr><td>Abbott Research, B.V.</td><td><b>0.7586</b></td></tr> <tr><td>Abbott Biotherapeutics Corp.</td><td><b>0.7647</b></td></tr> <tr><td>Abbott Research Center</td><td><b>0.7667</b></td></tr> <tr><td>Abbott Gmbh &amp; Co.</td><td><b>0.7692</b></td></tr> <tr><td>Abbott Gmbh &amp; Co Kg</td><td><b>0.7857</b></td></tr> <tr><td>Abbott BmbH &amp; Co. KG</td><td><b>0.7931</b></td></tr> <tr><td>Abbott GmbH &amp; Co. HG</td><td><b>0.7931</b></td></tr> <tr><td>Abbott Gmbh &amp; Ci, Kg</td><td><b>0.7931</b></td></tr> <tr><td>Abbott Gmbh &amp; Co. Dg</td><td><b>0.7931</b></td></tr> <tr><td>Abbott Gmgh &amp; Co. Kg</td><td><b>0.7931</b></td></tr> <tr><td>Abbott Products Gmbh</td><td><b>0.7931</b></td></tr> <tr><td>Abbott Gmbh &amp; Cco. Kg</td><td><b>0.8000</b></td></tr> <tr><td>Abbott Gmbh &amp;l Co. Kg</td><td><b>0.8000</b></td></tr> <tr><td>Abbott Biologicals B.V.</td><td><b>0.8065</b></td></tr> <tr><td>Abbot Gmbh &amp; Co. Kg</td><td><b>0.8276</b></td></tr> <tr><td>Abbott Healthcare Products, B.V.</td><td><b>0.8537</b></td></tr> </table>	assg_name	dist	Abbott Laboratories	0.0000	Abbott Biotherapeutics Corporation	<b>0.7429</b>	Abbott Gmbh & Co.Kg A Corporation	<b>0.7500</b>	Abbott Research, B.V.	<b>0.7586</b>	Abbott Biotherapeutics Corp.	<b>0.7647</b>	Abbott Research Center	<b>0.7667</b>	Abbott Gmbh & Co.	<b>0.7692</b>	Abbott Gmbh & Co Kg	<b>0.7857</b>	Abbott BmbH & Co. KG	<b>0.7931</b>	Abbott GmbH & Co. HG	<b>0.7931</b>	Abbott Gmbh & Ci, Kg	<b>0.7931</b>	Abbott Gmbh & Co. Dg	<b>0.7931</b>	Abbott Gmgh & Co. Kg	<b>0.7931</b>	Abbott Products Gmbh	<b>0.7931</b>	Abbott Gmbh & Cco. Kg	<b>0.8000</b>	Abbott Gmbh &l Co. Kg	<b>0.8000</b>	Abbott Biologicals B.V.	<b>0.8065</b>	Abbot Gmbh & Co. Kg	<b>0.8276</b>	Abbott Healthcare Products, B.V.	<b>0.8537</b>	
assg_name	dist																																								
Abbott Laboratories	0.0000																																								
Abbott Biotherapeutics Corporation	<b>0.7429</b>																																								
Abbott Gmbh & Co.Kg A Corporation	<b>0.7500</b>																																								
Abbott Research, B.V.	<b>0.7586</b>																																								
Abbott Biotherapeutics Corp.	<b>0.7647</b>																																								
Abbott Research Center	<b>0.7667</b>																																								
Abbott Gmbh & Co.	<b>0.7692</b>																																								
Abbott Gmbh & Co Kg	<b>0.7857</b>																																								
Abbott BmbH & Co. KG	<b>0.7931</b>																																								
Abbott GmbH & Co. HG	<b>0.7931</b>																																								
Abbott Gmbh & Ci, Kg	<b>0.7931</b>																																								
Abbott Gmbh & Co. Dg	<b>0.7931</b>																																								
Abbott Gmgh & Co. Kg	<b>0.7931</b>																																								
Abbott Products Gmbh	<b>0.7931</b>																																								
Abbott Gmbh & Cco. Kg	<b>0.8000</b>																																								
Abbott Gmbh &l Co. Kg	<b>0.8000</b>																																								
Abbott Biologicals B.V.	<b>0.8065</b>																																								
Abbot Gmbh & Co. Kg	<b>0.8276</b>																																								
Abbott Healthcare Products, B.V.	<b>0.8537</b>																																								

While the method we use is in principle very simple and transparent, the implementation requires detailed work and a lot of direct careful human input. One of the main difficulties is that internet search engines tend to return ‘too many result’, typically, the most popular domain names by internet users—that are not the companies’ domain names—appear at the top of the results lists returned by the search engines. These domain names belong to “Big internet platforms” such as facebook, linkedin, yahoo, bloomberg, wikipedia, reuters; “Online business registers” such as delawarelookup, delawarecompanies, registrycalifornia, eurofirmist; “Online newswires” such as nytimes, seattletimes, biznews, bostonglobe, businesswire; “Patent-related websites” such as patentquant, patentsobserver, trademarkia, patentbuddy; or “Investor-targeted websites” such as investopedia, investmentnews, investorroom, wikinvest. Overall, we have identified more than eleven hundred domain names that we exclude from the search engine results before performing the match on the domain name.

Manual work: Create domain “black list”
- Big internet platforms

facebook, linkedin, yahoo, bloomberg, wikipedia, reuters

- Online business registers

delawarelookup, delawarecompanies, registrycalifornia, euormlist

- Online newswires

nytimes, seattletimes, biznews, bostonglobe, businesswire

- Patent-related websites

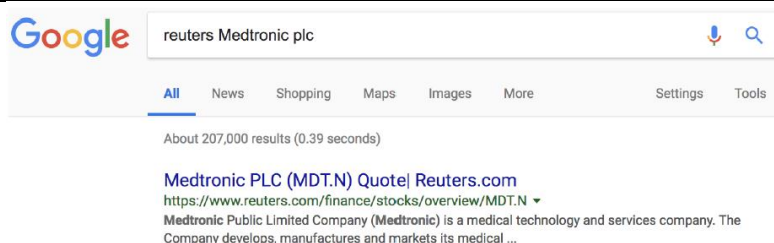
patentquant, patentsobserver, trademarkia, patentbuddy

- Investor-targeted websites

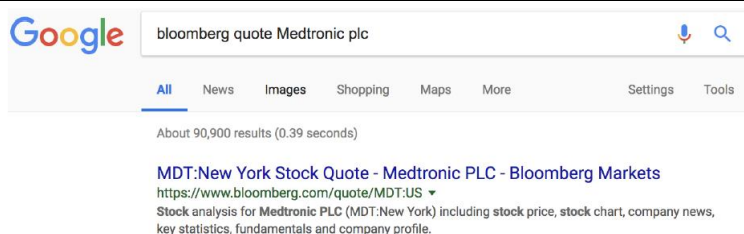
investopedia, investmentnews, investorroom, wikinvest

We augment the matching on domain names described above by using internet searches for stock market tickers of companies and we subsequently match on the tickers. To this end, we, for example, search for strings such as “reuters Medtronic plc”, “bloomberg quote Medtronic plc”, or “yahoo stock Medtronic plc” using [google.com](https://www.google.com). In all these cases, we obtain the result “MDT”, which correctly identifies the correct ticker of the company. We combine the matching on domain names with matching on tickers obtained from various sources, and create a set of indicator variables that describe whether a given patent assignee-company match has been confirmed using multiple different searches. Using these indicator variables, we can distinguish matches with different level of confidence in their validity.

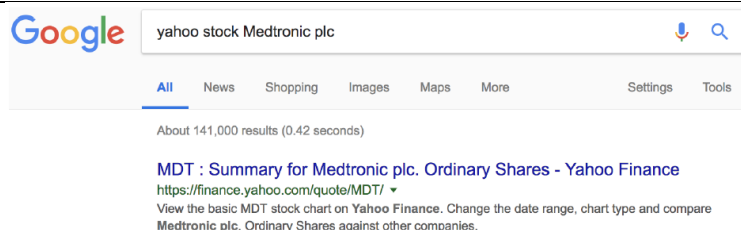
#### Use REUTERS search to collect ticker



#### Use BLOOMBERG search to collect ticker



#### Use YAHOO FINANCE search to collect ticker



We combine the matches obtained using internet searches with those obtained using fuzzy-string matching techniques and complement them using additional manual matching. Specifically, in terms of manual matching work: (i) We hand collect data on the most frequent not-matched assignees from the USPTO data. (ii) For multiple firms matched to the same domain name or ticker, we manually find the

correct company. In most of these cases, we assign the patent assignee strings to the parent company of the business structure the patent assignee belongs to. (iii) We compare our matching results to those provided by [Kogan, Papanikolaou, Seru, and Stoffman \(2017\)](#) and, for the period 2000-2010, we manually resolve all differences between our matches and those made publicly available in their [dataset](#). We resolve cases when the two datasets differ, that is, when a patent assignee is matched to a different company, as well as cases when one dataset indicates a patent assignee-to-company match while the other dataset does not.

While our method provides several advantages, the resulting dataset might not be suitable for all empirical applications. For example, when matching patents assigned to business group member firms, we effectively match such patents to the business group parent company irrespective of what legal entity inside the business group structure obtained the patent. As a result, for example, our data thus cannot be directly used to study the internal organization of innovation inside business groups. In short, while we identify the correct business group structure for each patent, a user of our data needs to make additional work to delineate where in the business group structure the patent belongs.

Furthermore, many subsidiaries of business groups originated through acquisitions. When searching for patent assignee strings of companies that became target firms in M&A transactions, we often obtain the domain name of the parent company of the business group that is the acquirer. In these cases, many target firms' patents end up being matched directly with the acquirers. Ultimately, after the acquisition, this is in fact the correct patent-parent company match, but the match is invalid before the M&A transaction occurs. As a result, our data thus, for example, cannot be directly used to study the in-house versus acquired innovation. In summary, while we identify the correct ultimate owner of the patent, a user of our data needs to make additional work to delineate the time when the subsidiary of the business group that was responsible for patenting was acquired.

Please, e-mail comments and suggestions that could be used to improve the data to [gcpd@arden.virginia.edu](mailto:gcpd@arden.virginia.edu).