Patentopia

A multi-stage patent extraction platform with disambiguation for certain semantic challenges

Belz, Andrea; Graddy-Reed, Alexandra; Shweta, FNU ; Giga, Aleksandar; Murali, Shivesh Meenakshi

DOI
10.1109/BigData55660.2022.10020918

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the 2022 IEEE International Conference on Big Data (Big Data)

Patentopia: A multi-stage patent extraction platform with disambiguation for certain semantic challenges

Andrea Belz
Viterbi School of Engineering
University of Southern California
Los Angeles, California USA
abelz@usc.edu

Alexandra Graddy-Reed
Sol Price School of Public Policy
University of Southern California
Los Angeles, California USA
graddyre@usc.edu

FNU Shweta
Viterbi School of Engineering
University of Southern California
Los Angeles, California USA
s779682@usc.edu

Aleksandar Giga
Faculty of Technology, Policy and Management
Delft University of Technology
Delft, the Netherlands
A.Giga@tudelft.nl

Shivesh Meenakshi Murali
Viterbi School of Engineering
University of Southern California
Los Angeles, California USA
smeenaks@usc.edu

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020918

***Abstract*—Bibliographic name disambiguation is an major semantic challenge, but critical to social sciences studies of important intellectual assets. Here we contribute to innovation research in several ways. We show a significant synonym problem in author names and discuss how a pre-processing heuristic step standardizing name variants helps, but homonyms generated with Chinese names are particularly difficult to resolve and manifest in an associated location list. Here we identify a new phenomenon of "onomastic profusion," the frequent use of certain words in firm names for semantic reasons that can confound disambiguation clustering algorithms. We illustrate these concerns with Patentopia, our customized platform accessing the PatentsView portal for the United States Patent and Trademark Office database and available for free academic use. This multi-stage system uses heuristics in concert with the PatentsView clustering process and reports meta-data to further assist analysis. As highly relevant use cases, we illustrate system performance with data derived from two important public innovation programs, I-Corps and Small Business Innovation Research (SBIR), and we close with implications for bibliometric analysis of current patent data.**

***Index Terms*—disambiguation, patents, NLP, bibliometric, SBIR, I-Corps**

INTRODUCTION

Research publications and patents represent important intellectual products for analysis of broad trends in knowledge creation. The associated field of digital library science is therefore important, but faces obstacles in semantic challenges generated by using author names, some of the most significant identifiers. Named entity (NE) disambiguation challenges include *synonyms*, in which a name appears in multiple forms due to name changes, presentation, or misspelling; and *homonyms*, wherein multiple entities share the same name [1].

These problems can exist for organizational names as well, but this challenge has not been recognized in bibliometrics scholarship. Onomastics, the study of proper names, has concerned itself mainly with names of people and places, and onomastic data scientists have directed their attention primarily to gender identification and author nationality [2]. However, the selection of a company name is a critical branding decision [3] and can signal firm quality [4]; therefore, firms choose their names carefully. Here we introduce the new disambiguation phenomenon of "onomastic profusion", in which certain words are commonly selected in organizational names for semantic reasons. In the invention context, words subject to this effect include "research," "advanced," "technology," and "development." This creates homonyms and thus confounds clustering disambiguation algorithms.

In this paper, we address disambiguation challenges for authors and assignees in the United States Patent and Trademark Office (USPTO) database. Several querying tools already exist, such as the simple Google Patents search interface; extension data sets, such as the Reliance on Science database linking patents back to academic papers [5]; and those studying historical patents [6]. Over the years, many methodologies have been developed using machine learning (ML) and natural language processing (NLP) [7]–[14], typically for fixed data sets that do not evolve in concert with the "raw" data. As a result, disambiguated databases currently in use (c.f. [12], [15], [16]), are not yet continuously updated. Fortunately, in the last few years, the USPTO has enabled broader access through its PatentsView platform [17] and associated Application Programming Interfaces (APIs) [18].

Here we report the development of an open academic access platform, Patentopia[1], for social scientists to easily access PatentsView APIs[2] and conduct additional disambiguation. The system searches either by an inventor name, such as a Principal Investigator (PI) of a federal research award; or

This research was funded in part by the National Science Foundation (NSF) I-Corps awards 1444080 and 1740721. Any opinions, findings, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the aforementioned organizations.

[1]https://sites.usc.edu/minerva/patentopia/
[2]https://patentsview.org/apis/api-endpoints/patents

assignee name, such as a firm or university. The time frame of these searches can be customized to speed processing time, as the entire USPTO database available through these APIs dates back to 1976; this wide range may not be necessary for some innovation studies. Patentopia enables both forward (outcome) and backward (explanatory variable) patent searches.

Patentopia supports large organizational- or inventor-level cross-sectional and panel studies with four key improvements to the currently available alternatives: (1) We identify errors in the PatentsView disambiguation process and provide metadata for potential mitigation strategies; (2) the system is synchronized with the PatentsView updates, providing nearly real-time information; (3) it enables efficient querying at scale by uploading a list of records in a comma-separated text format; and (4) the system is designed for intuitive use without the need for programming languages prior to integration with standard statistical packages such as R, Stata, or Python.

In this note, we discuss the specific disambiguation challenges that Patentopia is designed to address, including an explanation of onomastic profusion based in marketing theory. To illustrate inventor search, we use publicly named PIs of National Science Foundation (NSF) I-Corps awards, and for the assignee test case we use awardees of the National Aeronautics and Space Administration (NASA) Small Business Innovation Research (SBIR) program. We close with recommendations of further research to develop this bibliometric tool.

DISAMBIGUATION

A. Author names

Name disambiguation represents a significant obstacle to bibliometric studies; for instance, in one important academic database, two-thirds of the names were ambiguous and twenty percent had variant names, wherein an author records his/her name differently across multiple articles [19]. To resolve the inherent ambiguity in record linkage [20], three general approaches exist: Self-reporting identification, *a priori* generation of author identifiers, or fully automated processes [21]. This last option has naturally attracted extensive ML research along the traditional paradigms of supervised (classification), unsupervised (clustering), or semi-supervised (hybrid) models [22], depending on the use of labeled data sets.

However, the changing nature of the inventor pool creates new issues that have been less appreciated in the literature. First, the growing fraction of Chinese names poses challenges as they are typically shorter than Western names; family names are highly concentrated in this pool; and middle names are rare [1]. Thus, as Chinese names appear with greater frequency in databases, the homonym problem will dominate over the synonyms, creating a bias. Another problem is that women are more likely to have gender-neutral names [23], creating additional homonym opportunities and potentially confounding gender studies.

B. Organizational names

NE disambiguation approaches of company names have primarily relied on contextual clues, such as in short social media messages [24]. However, company names can be considered as brand names because they impart additional information [4]. For instance, companies with fluent (easy-to-pronounce) names have higher performance on several financial metrics [3], likely because these names convey familiarity.

Therefore, it is desirable to choose a name carefully. Two strategies are to create a name that is *descriptive* ("General Motors") or a *suggestive* ("Mr. Clean") [25]. An alternative path is to create or "*coin*" a new word ("Microsoft"), potentially attractive as it can be trademarked and can be used as a suggestive name [26]. A final possibility is an *arbitrary* word with limited reference to the product ("Camel" cigarettes).

While multiple taxonomies of brand names exist [26], [27], one simple way to categorize them is as *meaningful* (descriptive or suggestive) or *non-meaningful* (coined or arbitrary) [28]. This classification is consistent with research on brand names; for instance, meaningful names are easier to recall and generate better consumer response than non-meaningful ones [25], and brand names conveying a benefit lead to higher recall of that benefit [29].

In a search for familiarity, a low-cost approach is to select meaningful brand names. Viewed through the lens of competitive equilibrium theory, firms are likely to converge on the same words [30]. In the invention arena, this manifests as overuse of words such as "scientific," "research," "development," "advanced," or "technology" in brand names, leading to "onomastic profusion". Indeed, Klink noted that "using semantics can compromise the distinctiveness of the name" [31]. This semantic branding strategy has important consequences for disambiguation because it creates a new homonym problem similar in spirit to that of Chinese names [1] and confounding clustering.

In summary, organizational names act as brand names and are selected to convey certain characteristics. In the innovation arena, certain words related to invention are often chosen and embedded within the organizational name. This creates onomastic profusion wherein certain words linked to invention are over-represented, creating a new set of homonym problems previously unrecognized in bibliometric research.

C. Issues common to authors and organizations

As a final note, a source of synonym problems is common typographical errors, which include (in decreasing occurrence) omission, insertion, substitution and transposition errors [32]. Our experience is that in innovation databases, such as the public grant records discussed here, roughly 2% of the records face these issues, which are often discovered manually.

In principle, location data can be for further filtering [33], but in the United States several problems arise. For instance, the geography of innovation is distributed unequally throughout the United States [34], making associated location rules, such as state, difficult to implement evenly [35]. In other words, identifying an inventor uniquely in California is more challenging than in Wyoming and potentially creates a bias.

Therefore, multiple disambiguation processes must be managed simultaneously, with supervised or unsupervised learning

models possibly coupled to heuristic or rules-based identification [36], [37]. Patentopia addresses disambiguation problems with additional information, particularly fuzzy ratios (FRs). This metric varies with the length of the compared strings - for instance, the single-character substitution "Smith"/"Smyth" has a lower FR than "Anderson"/"Andersen". Patentopia employs a standard Python package[3] to evaluate differences between strings and reports these metrics in percentage points for both the inventor and assignee names, but does not make data selection decisions based on these values. The output is designed for integration with externally generated information, such as inventor curricula vitae or firm histories to inform further heuristics- or rules-based filtering.

System Architecture

PatentsView uses a clustering process to aggregate patents for both inventors and assignees [38], [39]. The Patentopia architecture proceeds along two different paths for the inventor and assignee search processes (Fig. 1). Module 1 represents pre-processing of the input data set. Module 2 directs the disambiguation process, with Module 2a requiring queries of a separate assignee database maintained by PatentsView in Module 3, unlike the simpler path of Module 2b for author searches. This architecture presents the following advantages: First, data updates are trivial since newer changes and synonym resolution for existing assignees can be simply applied as marginal changes to the existing database. This form of incremental updates is employed internally at PatentsView to generate disambiguation databases [38], [39]. Second, we can asynchronously update the database independently of other Patentopia development.

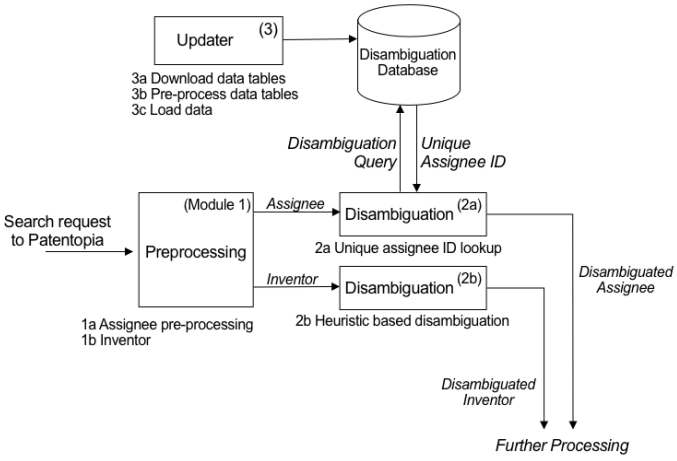

Fig. 1: Patentopia modules

Author Disambiguation: Inventors

Processing inventors

Patentopia proceeds as shown in Fig. 2. When a user uploads a list of inventors, Patentopia matches the last names exactly. However, the first name is matched only partially because of problems with nicknames (e.g., "Bobbie"/"Roberta"). Moreover, the name used in one record system may not match that of the USPTO database; we have seen this frequently in federal grant registries. As a result, a researcher can express the name twice - i.e., one variable listing "Bobbie Smith" and a second with "Roberta Smith". The second form of the name is optional and can be provided only for select observations, if desired. If provided, it is used for the Patentopia match process. We suggest trying Patentopia once with the initial list of names, then exploring a larger pre-processing effort based on the initial findings. Our experience is that a rule for conversions like "Bobbie" to "Roberta" has great value.

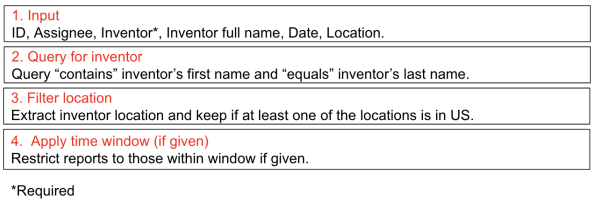

Fig. 2: Flowchart depicting the processing steps for inventor search.

Example challenges in inventor search

We report examples of name variations with select NSF I-Corps awardees (Table I). Patentopia reports both the inventor name and identification (ID) assigned by PatentsView. The inventor-ID relationship is a one-to-many match, in which highly productive inventors typically have many different identification numbers. We suspect that this may be due to submission from multiple organizations (i.e., different university employers or new research partners), changes in patent agents managing the filings, or other situational issues incidental to the actual content of the patents. These identification numbers can be altered dramatically with updates on PatentsView; indeed, the data structure may even change, such as a revision in the number of digits or a hyphenated labeling system. In general, we find that synonymous variations in Western name presentation (middle names, middle initials, or nicknames) cause the inventor search to run a greater risk of false negatives or missing patents. In contrast, faulty aggregation by the PatentsView system causes homonymous names, particularly of Asian descent, to have a high risk of false positives.

Patentopia reports all locations ever attributed to a single inventor ID to help track inventors who move - for instance, one who invents in graduate school and then moves to a faculty position at another university. As a result, while location is not required as an input field, it is highly recommended as it greatly eases the disambiguation process and is matched at the state level. We found that city matching was too restrictive because in large metropolitan areas, inventors are less likely to

[3]https://pypi.org/project/fuzzywuzzy/

TABLE I. Examples of disambiguation challenges for inventor search

Submitted		Returned from PatentsView by Patentopia			Correct?
Inventor Name	Location	Inventor Name	Inventor ID	Location(s)	
Ram Dantu	TX	Ram Dantu	2999395	Richardson, TX, US	Yes
Ram Dantu	TX	Ramanamurthy Dantu	272510	Richardson, TX, US	Yes
Shorya Awtar	MI	Shorya nmn Awtar	1168213	Ann Arbor, MI, US;Clifton Park, NY, US	Yes
Ben Wang	GA	Ben Wang	142565	Pingtung County, TW;Taichung, TW;Tallahassee, FL, US;Suzhou, CN;Atlanta, GA, US	Yes*
Ben Wang	GA	Benjamin Wang	567694	Carlsbad, CA, US;San Leandro, CA, US	No
Bin Yang	WA	Bin Yang	315403	Dongguan, CN;Hefei, CN;Pasadena, CA, US;Durham, NC, US;Xi'an, CN;Bloomington, IN, US;Wilmington, DE, US;Bala Cynwyd, PA, US;Waltham, MA, US;West Lebanon, NH, US;Hanover, NH, US;Richland, WA, US	Yes*
Bin Yang	WA	Bin Yang	127957	Singapore, SG;Zhuhai, CN;Shenzhen, CN;Guangdong, CN;Kunming, CN;Chengdu, CN;Shanghai, CN;Suzhou, CN;Wuxi, CN;San Diego, CA, US;Duluth, GA, US;Lanzhou, CN;San Carlos, CA, US;Columbus, OH, US;Dublin, OH, US;Beijing, CN;Bridgewater, NJ, US;Lincoln, NE, US;Mahwah, NJ, US;Fort Wayne, IN, US;Chappaqua, NY, US;Ossining, NY, US;Yorktown Heights, NY, US;Northborough, MA, US;Jilin, CN;Changchun, CN;Munich, DE;Herrenberg, DE;Stuttgart, DE	No

Inventor name is listed as given in public NSF award records. The Inventor ID is returned with PatentsView patent lists. The value for Correct is determined with manual search. An asterisk indicates that PatentsView has aggregated other patents with those of the inventor of interest, and thus the correctly attributed patents form a subset of all those returned by PatentsView. Table information was retrieved in August 2021.

live in the city of employment; however, Patentopia provides the city name for additional disambiguation support.

Table I illustrates some of the challenges in this process, using I-Corps awardees as the input name. Here we report the `inventor.key.ID` because it is a shorter identifier than the alternative `inventor.ID`, although this field may be deprecated in future versions of PatentsView. Patentopia reports a FR for inventor name but our experience suggests that it is difficult to create an accurate threshold value because there are so many differences in name presentations. Issues related to inconsistent inventor name registration include: (1) *Submission under multiple name spellings.* One example is inventors who submit their names differently, such as Ram Dantu, who has two distinct inventor IDs. In this case, the single location associated with the inventor contains a state matching that of the university assignee, and thus it can be used to confirm that these two IDs belong to the same inventor. Searching strictly for "Ram Dantu" would yield a missing observation. (2) *Submission with different middle name structures.* Shorya Awtar has an inventor ID associated with "no middle name" (nmn) reported as a middle name. An exact search without the inserted "nmn" string would sharply reduce the observed number of patents - but this inserted string appears to be used infrequently.

A concern with using PatentsView is the presence of inventor ID numbers incorrectly linked to disparate technologies and assignees. Examples (Table I) include: (1) *Modest misaggregation.* "Ben Wang" shows homonymous misidentification resolved with the location match, though one patent associated with this inventor number is a false positive. (2) *Large misaggregation.* "Bin Yang" returned additional names, such as "Bingrui Yang", "Junbing Yang," and others excluded from Table I for readability. The records shown in Table I still indicate the concern. The first record has some elements associated with the inventor of interest, whereas the second one does not; but homonymous observations remain in the first record.

To find PatentsView inventor IDs subject to large aggregation, we suggest statistical approaches; for instance, extracting the number of locations associated with the inventor and creating a threshold. We estimate that most inventors are associated with roughly five locations. Outlier records with with 20 or even 50 locations exist and can be carefully examined. In principle, assignee names can be used for matching but are subject to the onomastic profusion illustrated below.

NAMED ENTITIES: ASSIGNEES

Processing assignees

Disambiguation of assignee as the NE follows the same logic as the inventor analysis. Synonym problems exist, such as inconsistencies in institution labeling; for instance, the University of California Los Angeles may be referred to as "UCLA" in grant records and as the "Regents of the University of California" in patents.

The PatentsView hierarchical clustering algorithm [38] compares entity names with standard text string matching techniques and leverages other relevant information (inventors, locations, and patent classification). We developed a process to reverse some of the PatentsView clustering, by downloading the regularly updated `rawassignee` and `assignee` tables available on the PatentsView portal.

Our system executes the sequence of Fig. 3: It associates the user-submitted assignee name with a cluster ID using the `rawassignee` table, links the ID to its name from the `assignee` table, and then finds the assignee at issue for

each patent in the `rawassignee` table. In addition, we pre-process the `rawassignee` table to remove "LLC", "Inc.", etc. to align with the processing of input names. In effect, Patentopia uses the PatentsView cluster process to identify potentially associated assignee names, and then reverses the clustering. This extra step would be difficult for typical users of the PatentsView portal as it entails direct access to and processing of the raw data tables.

Example challenges in assignee search

To demonstrate patent extraction indexed by assignee search, we use a public list of NASA SBIR award recipients (Table II). With each patent Patentopia reports the FR score for the assignee name. For example, "Advanced Technologies Group, Inc." has a FR score of 50 with "Advanced Engineering Solutions, Ltd.," contrasting with 98 for the correct assignee.

A second matching score compares the inventor name, if provided, with all the patent's inventors and reports the highest FR. While Patentopia can execute a search without a inventor name, this variable clearly aids disambiguation. However, the case of Eltron Research & Development reveals poor inventor match scores (less than sixty percent), even for correctly assigned patents; however, in other cases (A&P Technology, Inc. and Advanced Technologies Group, Inc.) this score increases. This occurs because the award PI may not be the inventor or inventor disambiguation problems may exist.

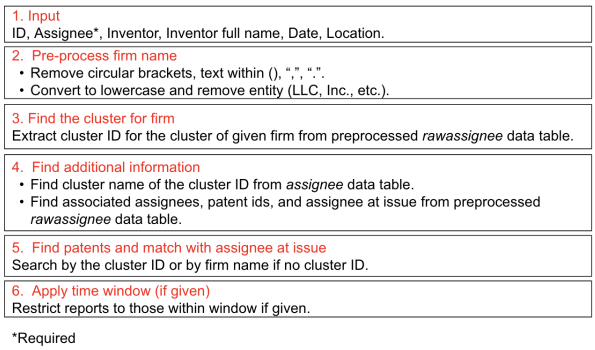

Fig. 3: Flowchart depicting the processing steps for search based on the assignee.

One challenge occurs because companies may not adhere to a standardized format in the patent application, confounding text matching metrics. For example, "Eltron Research & Development, Inc." on one patent may be listed as "Eltron Research, Inc." on another (Table II). In this case, the Boolean location variable can be used to correctly identify the relevant patents. The associated organization "Eltron Research, Inc." reveals relatively low FR scores for the assignee and inventor names (73% and 59%, respectively).

The onomastic profusion described above manifests in the associations created by PatentsView. For example, a cluster named "M&A Technology, Inc." also includes assignees "A&P Technology, Inc.", "G&H Technology, Inc.", and "R&H Technology, Inc.", even though manual inspection reveals them to be different companies. These are not isolated cases: A sample of 17 NASA SBIR winners - i.e., 17 clusters - revealed incorrect associations in six cases.

Indeed, searching the PatentsView API directly for "A&P Technology, Inc." or "Advanced Technologies Group, Inc" as assignee - the names recorded on the SBIR grants - yields no results because the API matches only the cluster name. Similarly, searching for "Advanced Engineering Solutions, Ltd." - the cluster name attached by the PatentsView to the actual grantee - returns more than 400 false positives. These discrepancies are sometimes identified on Google Patents, but this is infeasible at scale without a major programming effort.

The FR scores of names are useful but the threshold may not be obvious (Table II). Within the "M&A Technology, Inc." cluster, "A&P Technology, Inc." returns a value of only 82% for the match (because the term "Inc." was not present in the federal award record) but is the correct assignee at issue. On the other hand, "Advanced Technologies, Inc." is incorrectly associated with "Advanced Technologies Group, Inc." but shows a higher value for the FR at 88%. As a result, a FR threshold would likely need to be relatively low, with more rules to filter the data via the inventor name.

As location can be used for identification [33], we proceed with a low-fidelity version of prior work. For an assignee search, Patentopia extracts two different locations: (1) In the abridged report, the last known location of the *cluster* is reported; (2) In the complete report, each patent is labeled with the location of the *assignee at issue*. Point (1) is important because PatentsView attributes the last location of the assignee naming the cluster as the last location of the cluster. However, other associated assignees could have reported updated locations more recently. The location match in the case of Advanced Technologies Group, Inc. strongly suggests correct assignment; but these synonyms are found only through our de-clustering process (Fig 3). For this reason, when working with assignees in complex clusters, we recommend use of the locations in the complete file - i.e., at the patent level - for disambiguation. Otherwise, many of the cluster's patents are incorrectly associated with the wrong assignee, creating a high false positive rate. The onomastic profusion is such that correctly attributed patents for affected assignee names could easily represent only a few percent of the all patents identified for the cluster. In other words, the vast majority could be false positives.

FUTURE DIRECTIONS

Our methodology is designed to cast a wide net, with the idea that false positives can be identified and discarded, whereas false negatives (missing observations) are difficult to restore later. Of course room for improvement remains. Our current assignee search presumes that the inventor associated with the observation is named on the patent, but this does not

TABLE II. Examples of assignee clusters, Patentopia matching scores, and validation

Associated Assignee	Firm FR	Inventor FR	Location	Correct?
Eltron Research & Development Inc. (Eltron Research & Development, LLC)				
Eltron Research & Development, Inc.	99	37	TRUE	Yes
Eltron Research & Development, LLC	91	29	TRUE	Yes
Eltron Research, Inc.	73	59	TRUE	Yes
Eltron (London) Limited	39	34	FALSE	No
A&P Technology (M&A Technology, Inc.)				
A&P Technology, Inc.	82	88	TRUE	Yes
G&H Technology, Inc.	71	71	FALSE	No
M&A Technology, Inc.	71	17	FALSE	No
R&H Technology, Inc.	71	33	FALSE	No
Advanced Technologies Group Inc. (Advanced Engineering Solutions, Ltd.)				
Advanced Technologies Group, Inc.	98	88	TRUE	Yes
Advanced Technologies, Inc.	88	40	FALSE	No
Advanced Research and Technology Institute, Inc.	62	57	FALSE	No
Advanced Engineering Solutions, Ltd.	50	19	FALSE	No

Firm name is listed as given in SBIR records with the associated PatentsView cluster name noted in parentheses. Select assignees associated with the cluster are reported here. "FR" refers to fuzzy ratio scores reported in percent. Locations were compared with the SBIR records and matched on the state level. Table information was retrieved in August 2021. "Correct" indicates manual validation.

account for changing management teams. Network analysis identifying co-inventors could reduce false negatives.

It would also be useful to classify locations in Metropolitan Statistical Areas (MSAs) or other geographic indicators; for instance, an inventor living in Maryland but employed at a university in Washington, DC, could be identified through an MSA in a fashion not possible with our current location-match process. This would facilitate regional innovation studies.

Conclusion

Bibliometric analysis is an important capability to track production of intellectual assets, but it depends on accurate NE-record linkages. The relevance of such studies is greatly enhanced by leveraging dynamic databases rather than downloading a static corpus and deconstructing it. We leverage the tremendous effort of the PatentsView system and make several important contributions. First, we note the significant synonym problem may be partially alleviated with a pre-processing heuristic step to standardize name variants, but homonyms generated with Asian names are particularly difficult to resolve and manifest in an associated location list. Second, we describe onomastic profusion, explain why it appears, and how it creates previously unrecognized homonyms currently resolved incompletely by PatentsView. Finally, we show that location information can help resolve some disambiguations but creates potential biases due to innovation regional heterogeneity across the United States. Together with our novel platform, these findings assist innovation and policy scholars to integrate patent reports into their standard analysis flow.

I. Appendix

Submission

To use Patentopia, a researcher uploads a simple comma-separated value (csv) to the site. The interface is designed to be user-friendly (Fig. 4). Users can include up to six variables for each observation (row):

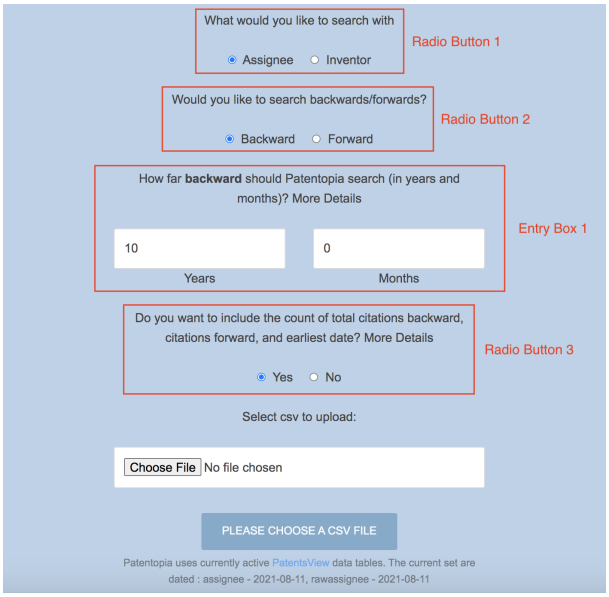

Fig. 4: Patentopia interface.

- Identification number, such as a federal award number or other unique identifier carried through the processing to assist in merging the patent information with the original data
- Inventor name
- Full or alternate version of inventor name
- Assignee name, typically a firm or university
- Location (US state)
- Observation-level date, such as for a survey response, award, or other time stamp.

The researcher indicates with Radio Button (RB) 1 if she is searching by inventor or assignee. Only that field is required in the researcher's csv file to execute a search to take place but the complementary field - that is, the inventor in the search by assignee or vice versa - and optional variables (e.g.,

location) are used for disambiguation. The inventor name can be provided in a second format, such as a full name with a middle initial as described further below, but this is optional. The user must remove individual extensions such as "Dr." and "Ph.D." for more accurate matching. However, Patentopia pre-processes entity identifiers (e.g., "Inc.", "LLC", "Corp."). Locations are matched at the US state level and reported as a Boolean TRUE/FALSE variable.

The time window option is important in studies of programs where an inventor or firm participates multiple times [40], making it critical to align data on the observational level. For instance, an inventor could be awarded an SBIR grant in 2011 and another in 2013, and it is desired to have a 5-year backward search on patent production. In the first case, the window would be 2006-2011; and in the second, 2008-2013. These two independent searches are linked to the appropriate observation via the SBIR award number.

A user selects backward or forward search relative to a date specified in the input file, as indicated with RB 2 and entry box 1 in the interface (Fig. 4), to enable use as an explanatory variable (backward) or as an outcome (forward). This temporal search window is entered in years and/or months format.

Citations and temporal filtering

Citations. Both backward and forward citation numbers are embedded within the PatentsView record. For a given patent (complete report), Patentopia returns the US patent numbers identified by PatentsView in both the backward and forward processes. In the abridged report, Patentopia sums the values as reported for a given inventor ID or cluster. Clearly the forward citation numbers would be expected to change over time as future patents cite an existing one.

Temporal filtering. In all searches, Patentopia can restrict the search in time and can limit reports to patents issued within the user-provided window (RB 3, Fig. 4). This can aid analysis: For example, a time window in backward search enables the researcher to limit patents to those more likely linked to the sampling date, which is important for prolific inventors with long histories. Similarly, restricting the temporal window for forward search increases the chance that the reported patents are associated with an intervention. In addition, this filter speeds processing.

If the input file does not contain a date, Patentopia returns the total patents associated with the inventor or assignee. If the user checks "yes" to the question of a total report, then the abridged report contains both the total citations and the citations within the time frame. In addition, "yes" causes Patentopia to return the first relevant patent date both in the database (dating back to 1976) and within the time frame. On the other hand, selecting "no" for this question forces Patentopia to report only the citations for the selected time frame, eliminating the variables related to "total" in the abridged report. If the time frame is implemented, the abridged summary reports as follows: The number of patents issued in the time frame is distinguished from that of the entire database, and similarly the back and forward citation counts. The complete report lists all patent numbers captured by those metrics for further investigation.

Output

Upon completion, Patentopia sends the user an email that repeats the submitted parameters. Output data exceeding 10,000 records are divided into subsets for individual distribution. A Patentopia email contains two attachments with patent data: a complete list of associated patents, and an abridged summary indexed by submitted assignee or inventor name. The system also creates two failure files that will be empty if all the searches, including those with no patents, were successfully returned. A search can potentially fail for reasons internal to PatentsView, networking issues, etc. To identify these cases, the two failure files list inventor/assignee names for failed searches and affected patent numbers. If a user receives output with entries in the failure files it is recommended to wait a day to resubmit the affected observations.

The output files are organized as the *complete* file listing each patent, and the *abridged* file consolidating the data by PatentsView inventor (assignee) ID in the search by inventor (assignee). The general architecture[4] is as follows:

Complete list. USPTO patent number and type; assignee name, location and type, list of inventors; filing and grant dates; backward and forward patent citations; and the United States Patent Classification (USPC) code. When FRs are evaluated, associated Boolean matching variables are set to TRUE for values of 100%. The location is matched as a Boolean TRUE on the state level. The output also contains a "full" matching variable that computes the logical AND of the name and location matches.

Abridged list. The patents of the complete list are aggregated at the identification level of the inventor or assignee, as appropriate for the input search parameter. Backward and forward citation counts are summed for patents within the time window specified; citation counts are also reported for all time. The FR for the name match and the logical test outputs from the complete list are reported here as well.

The PatentsView platform is updated regularly; for example, in the summer of 2021 the updates took place approximately monthly. In addition, the platform is constantly upgraded for improved performance (associated challenges are discussed further below). The dates of the PatentsView tables in use are shown on the portal. The information presented here was retrieved in August 2021.

ACKNOWLEDGMENT

A.B. served as co-PI and Research PI on the awards acknowledged herein prior to service at the National Science Foundation. To manage potential conflicts of interest she resigned from all roles associated with these awards and was recused from the associated NSF matters. The authors thank participants at the 2021 Academy of Management "New Data And Methods In Strategic Management Research"

[4]Details are available in the code books at https://sites.usc.edu/minerva/patentopia/info-for-scholars/

workshop, Matt Marx, Florenta Teodoris, and the Management of INnovation, Entrepreneurial Research, and Venture Analysis (MINERVA) lab.

REFERENCES

[1] D. Yin, K. Motohashi, and J. Dang, "Large-scale name disambiguation of chinese patent inventors (1985–2016)," *Scientometrics*, vol. 122, 11 2019.

[2] P. Roe, G. Lewison, and R. Webber, "The sex and ethnicity or national origins of researchers in astronomy and oncology in four countries, 2006-2007 and 2011-2012," *Scientometrics*, vol. 100, no. 1, pp. 287–296, 2014.

[3] T. C. Green and R. Jame, "Company name fluency, investor recognition, and firm value," *Journal of Financial Economics*, vol. 109, no. 3, pp. 813–834, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.jfineco.2013.04.007

[4] R. C. McDevitt, ""A" business by any other name: Firm name choice as a signal of firm quality," *Journal of Political Economy*, vol. 122, no. 4, pp. 909–944, 2014.

[5] M. Marx and A. Fuegi, "Reliance on science: Worldwide front-page patent citations to scientific articles," *Strategic Management Journal*, vol. 41, no. 9, pp. 1572–1594, 2020.

[6] M. J. Andrews, "Historical patent data: A practitioner's guide," *Journal of Economics and Management Strategy*, vol. 30, no. 2, pp. 368–397, 2021.

[7] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, vol. 50, no. 2, p. 104144, 2021. [Online]. Available: https://doi.org/10.1016/j.respol.2020.104144

[8] J. Raffo and S. Lhuillery, "How to play the "Names Game": Patent retrieval comparing different heuristics," *Research Policy*, vol. 38, no. 10, pp. 1617–1627, 2009.

[9] M. Trajtenberg, G. Shiff, and R. Melamed, "The "Names Game": Harnessing Inventors, Patent Data for Economic Research," *Annals of Economics and Statistics*, vol. 94, no. 93/94, p. 79, 2009.

[10] G. C. Li, R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, and F. Lee, "Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010)," *Research Policy*, vol. 43, no. 6, pp. 941–955, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.respol.2014.01.012

[11] M. Pezzoni, F. Lissoni, and G. Tarasconi, "How to kill inventors: testing the Massacrator© algorithm for inventor disambiguation," *Scientometrics*, vol. 101, no. 1, pp. 477–504, 2014.

[12] S. L. Ventura, R. Nugent, and E. R. Fuchs, "Seeing the non-stars: (Some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records," *Research Policy*, vol. 44, no. 9, pp. 1672–1701, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.respol.2014.12.010

[13] F. Teodoridis, J. Lu, and J. L. Furman, "Measuring the Direction of Innovation: Frontier Tools in Unassisted Machine Learning," 2021.

[14] A. Giga, A. P. Belz, R. J. Terrile, and F. Zapatero, "Helping the Little Guy: The Impact of Government Awards on Small Technology Firms," *Journal of Technology Transfer*, 2021.

[15] B. Hall, A. B. Jaffe, and M. Trajtenberg, "The NBER Patent Citations Data File: Lessons, Insights, and Methodological Tools," 2001.

[16] B. Balsmeier, M. Assaf, T. Chesebro, G. Fierro, K. Johnson, S. Johnson, G. C. Li, S. Lück, D. O'Reagan, B. Yeh, G. Zang, and L. Fleming, "Machine learning and natural language processing on the patent corpus: Data, tools, and new measures," *Journal of Economics and Management Strategy*, vol. 27, no. 3, pp. 535–553, 2018.

[17] N. Monath, A. Kobren, A. Krishnamurthy, M. R. Glass, and A. McCallum, "Scalable hierarchical clustering with tree grafting," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1438–1448. [Online]. Available: https://doi.org/10.1145/3292500.3330929

[18] USPTO, "USPTO APIs," 2021. [Online]. Available: https://developer.uspto.gov/api-catalog

[19] V. I. Torvik and N. R. Smalheiser, "Author name disambiguation in medline," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1552303.1552304

[20] H. Mihaljević and L. Santamaría, "Disambiguation of author entities in ads using supervised learning and graph theory methods," *Scientometrics*, vol. 126, no. 5, pp. 3893–3917, 2021.

[21] C. A. D'Angelo and N. J. van Eck, "Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation," *Scientometrics*, vol. 123, no. 2, pp. 883–907, 2020.

[22] H.-q. Han, Y. Yu, W. Lijun, X. Zhai, Y. Ran, and J. Han, "Disambiguating uspto inventor names with semantic fingerprinting and dbscan clustering," *The Electronic Library*, vol. 37, pp. 225–239, 04 2019.

[23] S. Lieberson, S. Dumais, and S. Baumann, "The instability of androgynous names: The symbolic maintenance of gender boundaries," *American Journal of Sociology*, vol. 105, pp. 1249–1287, 03 2000.

[24] N. Polat, "Experiments on company name disambiguation with supervised classification techniques," in *2013 International Conference on Electronics, Computer and Computation (ICECCO)*. IEEE, 2013, pp. 139–142.

[25] C. Kohli and R. Suri, "Brand names that work: A study of the effectiveness of different types of brand names," *Marketing Management Journal*, vol. 10, pp. 112–120, 2000.

[26] M. Danesi, "What's in a brand name? A note on the onomastics of brand naming," *Names*, vol. 59, no. 3, pp. 175–185, 2011.

[27] S. Arora, A. D. Kalro, and D. Sharma, "A comprehensive framework of brand name classification," *Journal of Brand Management*, vol. 22, no. 2, pp. 79–116, 2015.

[28] L. Todea, "Technology brands inspired by nature," in *Proceedings of ICONN 3*, 2015, pp. 856–867.

[29] K. L. Keller, S. E. Heckler, and M. J. Houston, "The Effects of Brand Name Suggestiveness on Advertising Recall," *Journal of Marketing*, vol. 62, no. 1, pp. 48–57, 1998.

[30] H. Hotelling, "Stability in Competition," *The Economic Journal*, vol. 39, no. 153, pp. 41–57, 1929.

[31] R. R. Klink, "Creating Meaningful New Brand Names: A Study of Semantics and Sound Symbolism," *Journal of Marketing Theory and Practice*, vol. 9, no. 2, pp. 27–34, 2001.

[32] J. J. Pollock and A. Zamora, "Collection and characterization of spelling errors in scientific and scholarly text," *Journal of the American Society for Information Science*, vol. 34, no. 1, pp. 51–58, 1983.

[33] G. Morrison, M. Riccaboni, and F. Pammolli, "Disambiguation of patent inventors and assignees using high-resolution geolocation data," *Scientific Data*, vol. 4, pp. 1–21, 2017.

[34] J. Guzman and S. Stern, "Where is Silicon Valley?" *Science*, vol. 347, no. 6222, pp. 606–609, 2015.

[35] E. Caron and N. J. van Eck, "Large scale author name disambiguation using rule-based scoring and clustering," in *Proceedings of the 19th international conference on science and technology indicators*. CWTS-Leiden University, Leiden, 2014, pp. 79–86.

[36] E. Iversen, M. Gulbrandsen, and A. Klitkou, "A baseline for the impact of academic patenting legislation in norway," *Scientometrics*, vol. 70, no. 2, pp. 393–414, 2007.

[37] C. A. D'Angelo, C. Giuffrida, and G. Abramo, "A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 257–269, 2011.

[38] N. Monath, C. Jones, and S. Madhavan, "PatentsView: Disambiguating Inventors, Assignees, and Locations," American Institutes for Research, Tech. Rep., 2020.

[39] N. Monath, S. Madhavan, C. DiPietro, A. McCallum, and C. Jones, "Disambiguating patent inventors, assignees, and their locations in patentsview."

[40] A. P. Belz, A. Graddy-Reed, I. Hanewicz, and R. J. Terrile, "Gender Differences in Peer Review of Innovation," *Strategic Entrepreneurship Journal*, 2022. [Online]. Available: 10.1002/sej.1429