Standardizing Patent Data Cleaning in a University Technology Transfer Office

M Srihari, Mary Mathew
Department of Management Studies, Indian Institute of science, Bangalore, India

Abstract--**There are multiple goals of a technology transfer office (TTO) based in a university system. Whilst commercialization is a critical goal, maintenance and cleaning of the TTO's database needs detailing. Literature in the area is scarce and only some researchers make reference to TTO data cleaning. During an attempt to understand the commercial strategy of a university TTO in Bangalore the challenge of data cleaning was encountered. This paper describes a case study of data cleaning at an Indian university based TTO. 382 patent records were analyzed in the study. The case study first describes the back ground of the university system. Second, the method to clean the data and the experiences encountered are highlighted. Insights drawn indicate that patent data cleaning in a TTO is a specialized area which needs attention. Overlooking this activity can have legal implications and may result in an inability to commercialize the patent. Two levels of patent data cleaning are discussed in this case study. Best practices of data cleaning in academic TTOs are discussed.**

I. INTRODUCTION

In recent years university systems have become conscious of the need to commercialize their research work, thus, engaging also in entrepreneurial activities [1]. The core activity that makes this happen is technology transfer of university research to other sectors, namely industry [2]. Technology transfer of novel research outputs resulting from the various internal labs of a university system ensures patenting their intellectual property and then licensing of these patents. This activity by which university research gets commercialized is carried out by the university Technology Transfer Office (TTO) [3] [4]. The TTO manages the patent database of universities and commercializes the patents, single or in portfolios. While much literature has emerged about the TTO and its activities, not much is written about cleaning a patent database in a university system. The focus of this paper is to dwell on patent data cleaning in order to make it ready for commercialization.

Literature involving university based TTOs [5], give insights into the operations of a TTO and on how patents are marketed by them. It shows that most TTOs market their patents via university websites and expect potential licensees to call on them for commercialization of the patents listed under the website. Also highlighted in this context is the "cradle-to-grave" approach for patent commercialization. This approach implies that highly skilled and dedicated staff own the complete patenting process right from the start of the invention disclosure until the point the patent is successfully commercialized. The success of commercialization of university patents is to some extent dependent on the diligence shown by the TTO staff in maintaining correct patent information in the university website. Human error in maintaining correct patent information can occur by overlooking important changes happening to the patent document over its life term. This can occur when the TTO staff is unskilled or under conditions of frequent staff turnover. Such errors can lead to the accumulation of incorrect patent information over time. Consequential errors occur because the staff entering and managing the patent database may not be the same staff that markets the database. So, TTOs need to ensure that the patent information in the university website is maintained correctly at all times by doing frequent checks. Failure to do so can affect commercialization prospects of these patents.

The study in [5] also indicates that most university TTOs keep internal records (not necessarily electronically) of their respective university patents in order to manage the technology transfer process. Normally this information involves data on all the different stages of a patent life cycle, including disclosure, patenting, licensing, maintenance and enforcement. At the stage when this internal patent data record is used by the TTO as the source for updating their marketing website then, special care needs to be taken to ensure that this internal record is first cleaned of incorrect patent information. This can be done by identifying the different types of mismatches and correcting the data before publishing it on any marketing website. We draw on insights from literature regarding this activity.

II. LITERATURE ON PATENT DATA CLEANING

There are many ways in which researchers have defined data cleaning. While data warehousing defines data cleaning in terms of data sanitizing using data masking techniques [6], others prefer to describe it in terms of finding of duplicates [7] and identifying inconsistencies in data [8]. One such definition refers to data cleaning as an activity by which data quality can be improved by means of detecting and removing errors and inconsistencies in data [9]. The history of data cleaning [10] mentions that data cleaning although having an equal possibility to influence the results of a study seemed to however receive very less attention from researchers. Since using incorrect data can lead to incorrect conclusions the data cleaning activity should ideally be the first step before embarking on any study involving data analytics irrespective of the field of application. An important component in the data cleaning process as explained in [11] consists of auditing the data and identifying discrepancies. Different methods of data cleaning spanning applications in different areas are discussed in [12], [13], [14], [15], [16] which illustrates the importance of the data cleaning in different fields.

When reviewing literature specific to patent data cleaning, it was found that most journal authors using patent data analytics explain details like - the sample data, the method used and the results and conclusions. Very little mention was found on pursuing patent data cleaning as an important and separate step. Patent data cleaning is explained in the nine-step tech mining process for the purpose of eliminating redundancies and unnecessary variations in patent and publications data [17]. The need for cleaning and standardizing patent data like inventor's address, city, state and country is explained in [18] and removing of duplicates and cleaning of key variables is explained [19] as part of patent data cleaning.

When reviewing literature specific to patent data cleaning in university based TTOs it was found that not much work has been done in this specific area. Patent data cleaning of a "raw list" of inventor patent data from a university TTO patent database, by comparing its accuracy with Delphion database, is explained in Sargossi and de la Potterie,2003 [20]. From this paper it can be understood that there seems to be a need to clean internally maintained patent data records in some TTOs by comparing them with other presumably cleaner databases.

III. METHODOLOGY

The case study method is adopted. The patent data cleaning levels and procedure used to clean internal patent records in the TTO of the Indian Institute of Science (IISc) is described in this section. The TTO of IISc is referred to as the Society for Innovation Development (SID). The procedure uses an important assumption that patent information available in the country's Patent Office (in this case India or the geography where the patent is filed) is the point of reference for the general public to evaluate patents. Anyone can assess features of the patent from the website of a Patent Office, for examples they can check for aspects like what is novel and protected, by whom, by how much and for how long. Hence, the Patent Office as a data source, which is usually the Patent Office website, is assumed to be the legally correct data source to follow. The following variables of patent data namely, Number *of Patent Records, Tracking Status, Renewal Information*, and *Assignee Information* can be obtained from these web sites.

The operational definition of the variables used in this paper is described below:

i. *Number of Patent Records:* Total count of patent documents which have been filed by the TTO in a given duration. Each patent document belonging to the same patent family is considered as a separate patent document for the count.

ii. *Tracking Status:* The current prosecution status of the patent document as tracked either at the TTO or at the Patent Office.

Tracking Status can take any one of the following values:

- *Patent under prosecution* – includes all patent applications which were filed at the Patent Office and are still under various stages of prosecution before the final patent grant.
- *Patent granted* – includes all patents which were successfully prosecuted and granted by the Patent Office and for which patent protection is active due to on-time payment of renewal fees.
- *Patent lapsed* – includes all patents which were under prosecution but were abandoned before final patent grant (or) includes all patents granted but for which patent protection has since ceased due to non renewal.

iii. *Renewal Information:* The state of renewal for a granted patent at the current time period.

iv. *Assignee Information:* The names of current assignees for a granted patent.

The different steps used for patent data cleaning of internal patent records in the TTO are explained below:

Step-1: Collect patent data independently for a given duration from the public website of the Patent Office by way of conventional patent search. The search can be done using the search field with name of the university as the Assignee Name.

Step-2: First level of patent data cleaning for *Number of Patent Records* and *Tracking Status*: Compare the entire list consisting of the patent data records internally available in the TTO with the list of all patent records collected from the website of the Patent Office obtained in Step-1. Identify mismatches and correct them by changing them as per the information in the website of the Patent Office to arrive at an interim list of cleaned patent data records.

Step-3: Second level of data cleaning for *Renewal Information*, and *Assignee Information*: Select granted patents from the interim list of patent data records obtained in Step-2. For each granted patent compare its renewal information and assignee information against corresponding details collected from the website of the Patent Office. Identify mismatches and correct them by changing them as per the information in the website of the Patent Office to arrive at the final list of cleaned patent data records suitable for commercialization.

IV. BACKGROUND OF THE TTO

The university chosen for this case study is India's premier research institute, the IISc, which is located in Bangalore city. It is one of the oldest academic centres in India and is well known nationally and internationally for its contribution to scientific growth and technology. The TTO of this university, the SID, was founded in the year 1991. One of the main goals of the TTO is to ensure that the innovations in science and technology from the university reach the outside world. To realize this goal the TTO interfaces with the university's Intellectual Property Cell (IPC). The main function of the IPC is to facilitate the filing and maintenance

of patents resulting from the research happening within the university. Since the IPC and the TTO together performed the role of patent data record keeping, management and commercialization the term TTO will refer to the roles the IPC plays too.

It was observed that the TTO maintained an in-house patent database which could be commercialized. This database was created to track the progress of all patent applications from the university right from the first filing at a Patent Office till the point the patent term expired. The process followed by the IPC staff for patent data record keeping and management of the TTO patent database was as follows:

1. A unique internal tracking number was assigned to each patent document filed at a Patent Office.
2. The first data record was created in the TTO patent database using this internal tracking number.
3. Updates to the patent document which resulted from the various stages of prosecution at the Patent Office such as tracking status, patent application number, patent grant number, patent grant date, patent renewal information, patent assignee information were subsequently appended to the patent record after looking up the corresponding internal tracking number which served as the primary lookup key in the patent database.

At the start of this study, there were a total of 229 patent data records in the TTO patent database starting from the year 1994 and ending in the year 2012. Table I shows the breakup of *Number of Patent Records* by geography i.e., India, US and Europe (other geographies were not included here), where the patent applications were filed for prosecution and by the *Tracking Status* of the patents as was maintained in the TTO patent database. Although other geographies are not mentioned in Table I, other than India, US and Europe, IISc has filed for patents in Australia, Brazil, Canada, China, Indonesia, Japan, Korea, Mexico, New Zealand, South Africa, Srilanka, Switzerland, Thailand and Vietnam.

TABLE I. *NUMBER OF PATENT RECORDS* FILED IN EACH GEOGRAPHY WITH CORRESPONDING *TRACKING STATUS* AS WAS MAINTAINED IN THE TTO PATENT DATABASE

***Tracking Status* as per TTO patent database**	***No. of Patent Records* by geography as per TTO patent database**		
	US	***Europe***	***India***
Patent under prosecution	43	0	96
Patent granted	28	2	44
Patent lapsed	6	7	3
Total	**77**	**9**	**143**
Grand total	**229**		

V. FINDINGS

When the different Patent Office websites were searched using patent search with the university name as Assignee Name and the filing duration as starting from the year 1994 till the end of year 2012, a total of 301 patent data records were obtained. Table II shows the details of the breakup of the *Number of Patent Records* by the Patent Office website from where the search results were obtained and by the *Tracking Status* of the patents as was observed in the Patent Office website.

TABLE II. *NUMBER OF PATENT RECORDS* AVAILABLE WITH CORRESPONDING *TRACKING STATUS* AS OBTAINED FROM SEARCHING THE PATENT OFFICE WEBSITE

***Tracking Status* as per Patent Office website**	***Number of Patent Records* available in the Patent Office website**		
	USPTO	***EPO***	***IPO***
Patent under prosecution	45	7	109
Patent granted	37	4	35
Patent lapsed	13	13	38
Total	**95**	**24**	**182**
Grand total	**301**		

This number, 301, did not match the 229 patent data records as was being maintained in the TTO patent database shown in Table I.

A. Findings from the first level of patent data cleaning for Number of Patent Records and Tracking Status:

1. Findings from comparing Number of Patent Records

It was found that 18 patent data records under US filings, 15 patent data records under Europe filings and 39 patent data records under Indian filings were missing in the TTO patent database when compared to the patent data records collected from the website of the Patent Office.

It was also found that 81 patent data records were missing in the Indian Patent Office website when compared to the patent data records in the TTO patent database. Table III gives the comparison results between the two patent data sources for the university.

TABLE III. COMPARISON OF *NUMBER OF PATENT RECORDS* BETWEEN TTO PATENT DATABASE VS PATENT OFFICE WEBSITE FOR EACH COUNTRY

***Tracking Status* of patent**	**US**		**Europe**		**India**	
	TTO database	***USPTO website***	***TTO database***	***EPO website***	***TTO database***	***IPO website***
Patent under prosecution	43	45	0	7	177	109
Patent granted	28	37	2	4	44	35
Patent lapsed	6	13	7	13	3	38
Missing patent records	**18**	**0**	**15**	**0**	**39**	**81**
Total	95	**95**	24	**24**	263	**263**
Grand total	**382**					

It was evident that there were 382 patent data records filed by university as against the 229 patent records being maintained in the TTO patent database and the 301 patent records which were obtained from searching the website of the different Patent Offices.

2. Findings from comparing Tracking Status

It was found that the status of the patents was being tracked incorrectly in the TTO patent database when compared to the status in the Patent Office website. Detailed comparison results of the status information for US, Europe

and India filings are shown in Table IV, Table V and Table VI, respectively.

TABLE IV. COMPARISON OF *TRACKING STATUS* OF PATENT DATA BETWEEN TTO PATENT DATABASE VS USPTO WEBSITE

Tracking Status as per USPTO website > ***Tracking Status as per TTO patent database***	Patent under prosecution	Patent granted	Patent lapsed	Missing patent records in USPTO website	**Total**
Patent under prosecution	35	5	3		**43**
Patent granted		28			**28**
Patent lapsed			6		**6**
Missing patent records in TTO patent data base	10	4	4		**18**
Total	**45**	**37**	**13**	**0**	**95**

TABLE V. COMPARISON OF *TRACKING STATUS* OF PATENT DATA BETWEEN TTO PATENT DATABASE VS EPO WEBSITE

EPO website status > ***TTO patent database status***	Patent under prosecution	Patent granted	Patent lapsed	Missing patent records in EPO website	**Total**
Patent under prosecution					**0**
Patent granted	1	1			**2**
Patent lapsed			7		**7**
Missing patent records in TTO patent data base	6	3	6		**15**
Total	**7**	**4**	**13**	**0**	**24**

TABLE VI. COMPARISON OF *TRACKING STATUS* OF PATENT DATA BETWEEN TTO PATENT DATABASE VS IPO WEBSITE

IPO website status > ***TTO patent database status***	Patent under prosecution	Patent granted	Patent lapsed	Missing patent records in IPO website	**Total**
Patent under prosecution	80	5	11	81	**177**
Patent granted		30	14		**44**
Patent lapsed			3		**3**
Missing patent records in TTO patent data base	29		10		**39**
Total	**109**	**35**	**38**		**263**

B. Findings from the second level of data cleaning for Renewal Information, and Assignee Information:

1. Findings from comparing the renewal information

Following observations were made when comparing the renewal information of individual granted patents between the TTO patent database and the website of the Patent Office:

Two patents granted in the US and one patent granted in Europe were within the six month period before their protection would cease due to non-payment of renewal fees.

Two patents granted in the US and three patents granted in India whose protection had already ceased were however in the revival period and could be restored by paying a penalty fee.

Both these important alerts which could help in extending the patent protection were not captured in the TTO patent database.

2. Findings from comparing the assignee information

When checking the assignee information of individual granted patents in the respective Patent Office websites it was found that there were four patents granted in the US, three patents granted in Europe and four patents granted in India which had other co-assignees in addition to the university. From among these it was observed that:

Two patents granted in the US and three patents granted in Europe were amongst the list of missing records from the TTO patent database.

Two patents granted in the US and four patents granted in India were present and co-assignee information was captured correctly in the TTO patent database.

This showed that some of the co-owned patents were captured in the database correctly while others were missed entirely, suggesting an inconsistent update.

C. Other process related findings

The above data mismatches were discussed with the IPC staff to understand the reason behind presence of inconsistencies in the data maintained in the TTO patent database. The reasons were found to be a few as described below:

The lack of prior knowledge on the kinds of data mismatches that can occur in internally maintained patent databases of a university TTO and understanding of standardized methods to identify and clean them being poor.

There was an assumption that granted patents in the TTO patent database were correct in all aspects as they were a result of successful prosecution at the Patent Office and hence, needed no further checks and updates over their remaining life term.

The updates happening on the patent records at the Patent Office were not automatically fed into the TTO patent database but were being manually keyed in by the TTO staff.

Some of the Indian filed patents which were missing in the Indian Patent Office website could be because of delayed data upload on the Indian Patent Office website by its systems staff.

VI. DISCUSSION

From the case study it was clear that there were gaps in the process followed by the IPC staff to update the patent records for the TTO patent database. This resulted in the following types of mismatches - 1) missing patent data records 2) incorrect status information 3) missing renewal information and 4) missing assignee information.

To avoid mismatches due to missing patent data records and incorrect status information it is recommended to incorporate frequent checks in the TTO patent database by comparing the number and status of patent data records with the data available in the website of the relevant Patent Office. This will help ensure that all patenting strategies and commercialization decisions taken by the TTO are done on the full and complete set of patent data available in the TTO patent database.

It is imperative that TTOs involved in commercialization of patents maintain correct renewal and assignee information at all times for the patents being commercialized. Renewal

information is essential to update marketing staff on whether the patent protection is legally active. Incorrect renewal information can lead to the assumption that a patent whose protection had ceased due to non-payment of maintenance fee is still active. Similarly, after a patent is granted it may undergo a transfer of ownership through assignment [21]. This is done by changing the assignee details on the patent. Transfer of ownership is transfer of patent rights in part or in full. There can also be cases where a patent was jointly applied by the university in collaboration with other organizations (private companies or Government firms), in which case, it will have multiple owners as assignees. Incorrect assignee information can lead to the assumption that a patent which had undergone full transfer of ownership or which was owned partly or was applied jointly is still fully owned (not co-owned) by the university. This has great implications on commercialization practice. Such incorrect details can have legal implications arising from trying to publish, market or license out invalid patents or co-owned patents. Litigations arising out of such problems can have an adverse effect on a TTOs licensing prospects [22].

To avoid this it is recommended that the renewal and assignee information of granted patents be compared with the information available in the website of the relevant Patent Office and cleaned at least once (ideally, periodically) before embarking on any commercialization activity. For renewals, since Patent Office websites [23] have already pre-defined maintenance fee and payment intervals, a good practice is to proactively identify the payment calendar for each patent and configure alerts based on this calendar.

VII. CONCLUSION

University based TTOs in developing countries such as India, can find themselves in problem situations if they seem to have a practice of using only their in-house TTO patent database as a reference while embarking on patent commercialization activities. This database is used to capture the patent information emerging from academic research right from the time of invention disclosure until the time the patent term expires. There seems to be a tendency in the TTO to assume that the patent details present in the database are correctly maintained and that there may not be a big need to clean it at frequent intervals. But these assumptions can go wrong over time due to human errors. As seen in this case study there could be patent data records which are missing completely or which could have incorrect status in the database. This can lead to the TTO overlooking them in its various strategies and commercialization studies. The renewal information of a granted patent will change over its life term similarly assignee information can change after a patent is granted. Hence, these aspects need to be checked regularly for mismatches against a standard public database like the website of the Patent Offices in which the university files its patents. These practices will help the TTO ensure that all its decisions are made on full and complete data present in the TTO patent database. It will also help avoid any legal problems that can arise out of using incorrect renewal status and ownership information while commercializing protected technology.

REFERENCES

[1] H. Etzkowitz, "Research groups as 'quasi-firms': the invention of the entrepreneurial university," *Res. Policy*, vol. 32, no. 1, pp. 109–121, Jan. 2003.

[2] T. R. Anderson, T. U. Daim, and F. F. Lavoie, "Measuring the efficiency of university technology transfer," *Technovation*, vol. 27, pp. 306–318, 2007.

[3] J. Sandelin, "University technology transfer in the US: History, status and trends.," *Beijing Int. forum Strateg. IURP.*, 2002.

[4] N. Baldini, R. Grimaldi, and M. Sobrero, "Institutional changes and the commercialization of academic knowledge : A study of Italian universities ' patenting activities between 1965 and 2002," *Res. Policy*, vol. 35, pp. 518–532, 2006.

[5] M. F. Allan, "A Review of Best Practices in University Technology Licensing Offices," *J. Assoc. Univ. Technol. Manag.*, vol. 13, no. 1, pp. 57–69, 2001.

[6] S. E. Fienberg, "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation," *Stat. Sci.*, vol. 21, no. 2, pp. 143–154, May 2006.

[7] W. E. Winkler, "Data Cleaning Methods," *Proc ACM SIGKDD Work. Data Cleaning, Rec. Linkage, Object Consol.*, 2003.

[8] H. Yu, Z. Xiao-yi, Y. Zhen, and J. Guo-quan, "A universal data cleaning framework based on user model," *Comput. Commun. Control. Manag. 2009. CCCM 2009. ISECS Int. Colloq.*, vol. 2, no. IEEE, pp. 200–202, Aug. 2009.

[9] E. Rahm and H. H. Do, "Data Cleaning : Problems and Current Approaches," *IEEE Data Eng. Bull.*, vol. 23, no. 4, pp. 3–13, 2000.

[10] J. Van Den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, "Data Cleaning : Detecting , Diagnosing , and Editing Data Abnormalities," *PLoS Med.*, vol. 2, no. 10, pp. 966–970, 2005.

[11] V. Raman and J. M. Hellerstein, "Potter ' s Wheel : An Interactive Data Cleaning System," *VLDB*, vol. 1, pp. 381–390, 2001.

[12] T. T. Aye, "Web log cleaning for mining of web usage patterns," *Comput. Res. Dev. (ICCRD), 2011 3rd Int. Conf.*, vol. 2, no. IEEE, pp. 490–494, Mar. 2011.

[13] W. Lup Low, M. Li Lee, and T. Wang Ling, "A knowledge-based approach for duplicate elimination in data cleaning," *Inf. Syst.*, vol. 26, no. 8, pp. 585–606, Dec. 2001.

[14] J. W. Osborne, "Data Cleaning Basics: Best Practices in Dealing with Extreme Scores," *Newborn Infant Nurs. Rev.*, vol. 10, no. 1, pp. 37–43, Mar. 2010.

[15] J. M. Hellerstein, "Quantitative Data Cleaning for Large Databases," *United Nations Econ. Comm. Eur.*, 2008.

[16] A. Loureiro, L. Torgo, and C. Soares, "Outlier Detection Using Clustering Methods : a data cleaning application," *Proc. KDNet Symp. Knowledge-based Syst. Public Sect.*, 2004.

[17] A. L. Porter and S. W. Cunningham, *TECH MINING Exploiting New Technologies for Competitive advantage*. 2005, pp. 324–326.

[18] M. Trajtenberg, G. Shiff, and R. Melamed, "The 'names game': Harnessing inventors' patent data for economic research," *Natl. Bur. Econ. Res.*, no. w12479, 2006.

[19] B. H. Hall, A. B. Jaffe, and M. Trajtenberg, "Market value and patent citations: A first look," *No. w7741. Natl. Bur. Econ. Res.*, 2000.

[20] S. Saragossi and B. van P. de la Potterie, "What Patent Data Reveal about Universities : The Case of Belgium," *J. Technol. Transf.*, vol. 1, no. 28, pp. 47–51, 2003.

[21] USPTO.gov: "Change Ownership", Retrieved 30/01/2014, World Wide Web, *http://www.uspto.gov/patents/process/changeownership.jsp*

[22] S. Shane and D. Somaya, "The effects of patent litigation on university licensing efforts," *J. Econ. Behav. Organ.*, vol. 63, pp. 739–755, 2007.

[23] USPTO.gov: "Maintain Your Patent", Retrieved 30/01/2014, World Wide Web, *http://www.uspto.gov/patents/process/maintain.jsp*