DIP-BR: An Open and Network-Based Dataset of Brazilian Patents

Pablo Vasconcelos da Cruz[1], **Iuri V. A. Lima**[1], **Danilo B. Seufitelli**[1], **Daniel H. Dalip**[1], **Fabrício P. V. de Campos**[2]

[1]Centro Federal de Educação Tecnológica de Minas Gerais
Belo Horizonte, MG – Brasil

[2]Universidade Federal de Juiz de Fora
Juiz de Fora, MG – Brasil

{iuriveraslima, pablo.vasconcelos.cefet}@gmail.com,
{hasan, danilobs}@cefetmg.br, fabricio.campos@ufjf.edu.br

Abstract. *The gap between research institutions and industry leads to a significant challenge for technology transfer. Patent analysis is a strategic tool for unveiling collaborations between companies and research institutions. In this context, we introduce DIP-BR, an open research dataset of Brazilian intellectual property. This dataset is enriched by applying a deduplication algorithm to standardize names and using machine learning techniques to classify each patent holder as a research institution, company, or individual. The primary contribution of this work is the resulting enriched dataset, which is structured through network modeling and clustering. DIP-BR serves as a tool for analyzing trends and visualizing the dynamics of innovation in Brazil.*

1. Introduction

Technology transfer is challenging for Science and Technology Institutions (ICTs) due to the gap between institutions conducting technological research and companies interested in new market-scale solutions [Fujino and Stal 2007]. Co-owned patents between companies and ICTs are an indicator of technology transfer and demonstrate their collaboration. These patents grant exclusive rights for both parties, strengthening their market position.

In this context, patent analysis evaluates technological innovations and is a strategic tool for planning, trend monitoring, and identifying innovation opportunities [Trappey et al. 2011, Asitah et al. 2024]. The Brazilian government[1] has recognized this relevance and promoted initiatives to expedite the granting of patents and trademarks in Brazil to strengthen the national innovation environment.

In [Rezende et al. 2023] we presented a patent dataset, containing information on patent requirements from June 2018, and a process that automates the retrieval and processing of these publications. We crawled the patent bulletin[2] from the National Institute of Industrial Property (INPI, an acronym for Instituto Nacional da Propriedade Industrial in Portuguese). This institute regulates the granting of industrial property rights in Brazil.

[1]"INPI e MDIC avançam para agilizar patentes e marcas no Brasil". Available in `https://bit.ly/InpiMdic`. Accessed in 2025 March, 27.

[2]`https://revistas.inpi.gov.br/rpi/`

The INPI bulletin provides a weekly update, offering patent holders information on newly filed patent applications, related requirements, and updates regarding Brazilian patents.

This dataset includes information such as the patent holder, patent title, category, and other relevant data. However, the name of the patent owner is stored as plain text, leading to inconsistencies, with different variations potentially referring to the same organization. To address these limitations, we introduce DIP-BR (***D**ataset of **I**ntellectual **P**roperty in **BR**azil*), an open dataset that enable the analysis of relationships between companies and ICTs in Brazil. As a result, we collected all patents published in the INPI bulletin since 2018. The database comprises $256,909$ patents and $67,869$ owners. We enriched such dataset by applying a deduplication algorithm to standardize names and using machine learning techniques to classify each patent holder (research institution, company, or individual). The final dataset is available at `https://doi.org/10.5281/zenodo.15757353`. This work's primary contribution is the enriched dataset structured with network modeling and clustering.

The key motivation for this study is the need to promote more effective integration among the various agents that comprise the innovation ecosystem. By presenting a dataset that maps existing collaborations between patent holders, this research seeks to facilitate new studies that want to analyze the connection between academic institutions, research centers, and companies. This convergence enhances technology transfer and knowledge exchange, fostering the development of innovative solutions that can positively impact society and the economy.

The remainder of this paper is structured as follows: Section 2 reviews related works, highlighting existing gaps in the patent landscape. Section 3 describes the process to collect patent data from the INPI and presents the results of an exploratory data analysis to uncover initial insights. Section 4 details the methodology, including data preprocessing, deduplication techniques, network construction, and classification approaches. Section 5 discusses the experimental findings, highlighting key patterns in patent collaboration and technological development. Finally, Section 6 summarizes our main contributions, outlines potential future directions, and addresses the study's limitations.

2. Related Work

The study of patents has become essential for various organizations, whether from a legal or innovation management perspective. They seek to identify technological trends, assess the potential of promising patents, and monitor strategic competitors [Abbas et al. 2014]. There are different approaches to patent analysis; many focus on predicting emerging technologies, a fundamental stage in the research and development process. For example, [Kim and Bae 2017] proposed a method based on forming technology clusters using cooperative patent classifications. These clusters are analyzed regarding technological combinations, and various indices are applied to assess their innovative potential.

Patent analysis can also be used to visualize the effort invested in developing and advancing specific types of technology. Using an international patent database, [da Silveira et al. 2021] identified the main actors behind technologies aimed at controlling air emissions from agricultural machinery in Brazil. Such a type of study requires a well-organized and enriched patent dataset containing relevant information. In 2023, [Suzgun et al. 2023] introduced a large-scale dataset with over 4.5 million patents. How-

Table 1. Technological areas of the International Patent Classification (IPC) and their respective codes.

Code	Technological Area
A	Human Necessities
B	Performing Operations; Transporting
C	Chemistry; Metallurgy
D	Textiles; Paper
E	Fixed Constructions
F	Mechanical Engineering; Lighting; Heating; Weapons; Blasting
G	Physics
H	Electricity

Table 2. Examples of duplicates for the UFOP institution and its associated patents for each record.

Institution Name	# of Patents
Universidade Federal de Ouro Preto	100
UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP	53
UNIVERSIDADE FEDERAL DE OURO PRETO ? UFOP	2
UNIVERSIDADE FEDERAL DE OURO PRETO UFOP	1
UNIVERSIDADE FEDERAL DE OURO PRETO- UFOP	1
UNIVERSIDADE FEDERAL DE OURO PRETO - UFOP	1

ever, it includes only patents filed in the United States. Such a limitation does not represent all innovation types and cannot be used in technology development for other markets.

In Brazil, technological output studies concerning patent applications are still limited and often focus on specific knowledge areas. As an example regarding the Biotech area, [Costa et al. 2018] analyzed the biotechnology sector in northeastern Brazil and identified a significant increase in technological production starting in 2009. Indeed, those studies rely on domain-specific datasets. In contrast, our dataset enables a broader and more systematic analysis of Brazil's patent landscape. By offering an enriched dataset with disambiguated and categorized patent holders (companies, individuals, or ICTs), our data allows large-scale studies across multiple sectors. Additionally, the clustering and deduplication processes enhance the reliability of institutional mappings, allowing for more accurate identification of collaboration patterns and innovation trends.

3. Brazilian Patent Data: Extraction, Processing, and Analysis

A patent of an invention is an exclusive title granted to the knowledge holder, who may be an individual, a company, or a science and technology institution (ICT). Furthermore, co-ownership occurs when a patent has more than one holder. Hence, all holders can exploit the patented object if they obtain reciprocal authorization [Matias et al. 2020].

The National Institute of Industrial Property (INPI) is the government agency responsible for granting industrial property rights in Brazil. The INPI bulletin is published weekly and often updates the technological applications submitted in Brazil. Such publications in the INPI bulletin include granted patents and those still under review. Hence, we build a Python script to access the bulletins in XML format from the agency's official website and extract all patent records with published decisions.

The crawled data comprises the patent title, owners and their location, the patent demand the owners need to solve, and the International Patent Classification (IPC). The IPC system classifies patents according to the different technological areas they belong to [de Almeida Chaves et al. 2019]. The classification code is typically quite specific. Here, we focus on the first character that represents the technological area of the patent, as shown in Table 1. Note that one patent can receive multiple codes, including codes from different classes, to represent all the technological areas it covers. Next, Section 3.1 presents the data-cleansing process and Section 3.2 an exploratory data analysis.

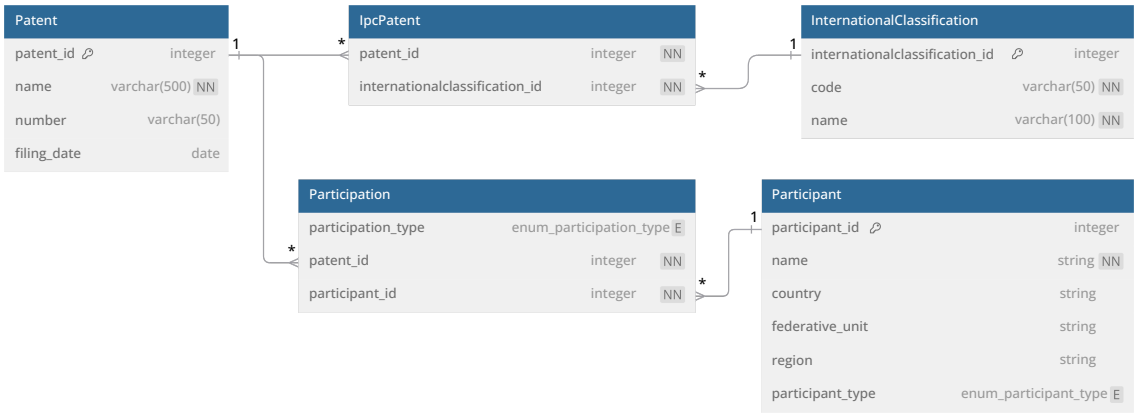

Figure 1. Entity Relationship Diagram

3.1. Data Cleansing

The INPI bulletin provides semi-structured data, meaning that while the information follows a specific format, it lacks strict consistency and standardization. As a result, data entry inconsistencies frequently introduce variations in entity names. Different records may list the same patent holder under multiple name formats due to inconsistencies in spelling, abbreviations, punctuation, or typographical errors. Additionally, data contributors may include or omit acronyms and special characters, or rearrange words when recording institutional names. The lack of unique patent holder identification difficult to map collaborations and analyze relationships of the innovation ecosystem. Without proper deduplication, multiple records for the same entity could skew results, leading to an inaccurate representation of patent counts for institutions, researchers, or companies.

One of the most frequent inconsistencies in institutional names arises from variations in acronyms and separator characters. Table 2 illustrates an example of data duplication, showing different ways of writing ***UFOP*** and the number of patents associated with each variation. To address this issue, we applied an algorithm that separates the name and the acronym using the most commonly used separator characters. Then, we compare the name with the others already found. This approach reduced most of the duplicate records. However, less frequent inconsistencies (e.g, typographical errors and more complex variations in institutional names) were manually reviewed and corrected. The Methodology Section further details this process and comprehensively explains the techniques applied.

3.2. Exploratory Data Analysis

After extracting and cleaning the data, we identified $256,909$ patents and $67,869$ patent holders. Notably, $20,945$ of these patents were associated with multiple holders, involving a total of 18,634 distinct holders in collaborative efforts. With this refined dataset of collaborative patents, available at `https://doi.org/10.5281/zenodo.15757353`, we present an exploratory analysis focusing on their key characteristics: the patent holder types, their countries and states of origin, and their IPC.

Figure 1 illustrates a simplified version of this paper's entity-relationship diagram. Table 3 provides the data dictionary, detailing the patents, holders, and their collaborations, while Table 4 summarizes the number of instances collected for each entity. The

Table 3. Data dictionary for the patent dataset

Field	Type	Description
Patent Holder		
holder_id	int	Patent holder identifier
fullName	string	Full name of the holder (normalized)
country	string	Country code of the holder
state	string	If Brazilian, indicates the holder's state
region	string	If Brazilian, indicates the holder's region
type	string	Applicant classification: Person, Company, or Research Institution
cluster_id	int	Cluster id of the holder
Patent Holder Collaboration Network		
holder_u_id	int	The ID of the holder u
holder_v_id	int	The ID of the holder v
jointly_filed_patents	int	The total number of patents that holder u and v have filed together.
jointly_granted_patents	int	The total number of patents that holder u and v have been granted together.
jointly_granted_patents_per_ipc	int[]	Granted patents, broken down by IPC classification, that the holders u and v have been granted together.
jointly_filed_patents_per_ipc	int[]	The number of patents jointly filed by holders u and v, broken down by IPC classification.
Patent		
applicationNumber	string	Patent application number
filingDate	string	Date the application was filed
grantDate	string	Date the patent was granted
nationalPhaseDate	string	Date of entry into the national phase
internationalApplication.pctNumber	string	PCT application number
internationalApplication.pctDate	string	Filing date of the PCT application
internationalPublication.wipoNumber	string	WIPO publication number
internationalPublication.wipoDate	string	Date of WIPO publication
title	string	Patent title
ipcCodes	string[]	List of IPC classification codes assigned to the patent
holders	int[]	List of patent holder ids
inventors	string[]	List of inventor names
priorityClaims	object[]	List of union priority claims
priorityClaims.countryCode	string	Country code of the priority claim
priorityClaims.priorityNumber	string	Priority claim number
priorityClaims.priorityDate	string	Date of the union priority claim
divisionalApplication.filingDate	string	Filing date of the divisional application
divisionalApplication.applicationNumber	string	Number of the divided application
parentApplication.filingDate	string	Filing date of the parent application
parentApplication.applicationNumber	string	Number of the parent application
events	object[]	List of legal events related to the application
events.code	string	Legal event code
events.description	string	Description of the legal event
events.bulletinNumber	number	Bulletin number where the event was published

Table 4. Dataset Entities Overview

Entity	Description	Count
Patents	Patent instances collected.	256,909
Holders	Unique patent holders	67,869
Collaborations	Co-ownership relationships (edges) between holders.	23,118
IPC Codes	Unique International Patent Classification codes.	48,586

primary contributions of this work are the preprocessing methodology for patent holders and the construction of their collaboration network.

Specifically in Brazil, Figure 2 classifies the number of co-holders by Brazilian state, allowing us to visualize the states and regions with the highest participation in collaborative patents. São Paulo ranks first with 36% of the co-holders, followed by Minas Gerais with 12%. The Southeast and South regions account for the largest share, while the Central-West and Northeast regions have slightly lower participation. The North region has the lowest co-holders, representing only 1.83%.

4. Methodology

Understanding the collaborative dynamics in patent filings requires a methodology that extracts, refines, and analyzes data to reveal meaningful patterns. A key aspect involves handling inconsistencies in patent holder names and developing analytical techniques to classify and interpret the entities engaged in innovation. Furthermore, we propose model-

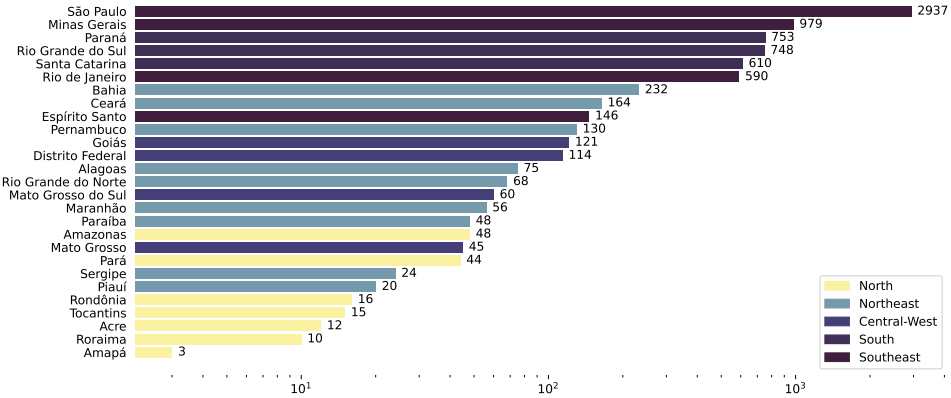

Figure 2. Number of Patent Holders by Brazilian State.

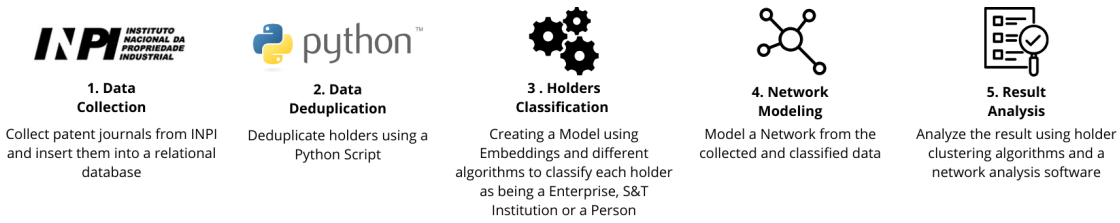

Figure 3. The methodology overview details the sequential steps from data collection to network-based collaboration analysis.

ing the relationships among patent holders as a collaboration network to comprehend the structure of innovation networks.

Hence, we propose a methodology composed of five main steps: (1) data collection from the INPI journals; (2) data deduplication; (3) classification; and (4) result analysis. Each step ensures that the extracted information accurately represents the relationships between individuals, institutions, and companies in the innovation ecosystem. The collaboration network derived from this process allows us to analyze connectivity patterns, identify key actors, and assess the overall structure of technological cooperation in Brazil. Figure 3 illustrates an overview of the methodology, outlining the sequential stages of the process and highlighting the integration of network modeling into the analysis. Next, we detail each methodology step.

Data Collection. The weekly patent bulletins are available by INPI in XML format through a download link.[3] We downloaded all bulletins from June 2018 to February 2025 and stored them in a relational database. Although the dataset only includes patents filed in Brazil, each record contains the holder's name, its international classifications, and country of origin. For Brazilian holders, the record also includes the corresponding state of origin.

Data Deduplication. Due to how INPI provides the data, some patent holders appear under different names across multiple records, making deduplication an essential part of the methodology. We identified a recurring pattern in the variations of the same institution's name. These differences usually stem from a separator character and the presence

[3]The bulletins are available at `https://revistas.inpi.gov.br/rpi/`

of an acronym after the name. We developed an algorithm that leverages the most common discrepancy in holder names, which is typically the presence of a symbol and the organization's acronym. While this approach cannot capture more subtle variations, such as typographical errors or missing words, it is sufficient for handling most cases.

The deduplication algorithm iterates through all holders, cleaning the records by removing duplicate spaces and converting all letters to uppercase. Then, it splits the record into a name and an acronym using the most frequent separator symbols. From these two components, the algorithm groups holders with the same name. This deduplication process reduced the number of distinct holders from $73,047$ to $67,869$.

Holders Classification. One of the objectives of this study is to analyze the relationships between different types of holders and their characteristics. Since the collected data does not provide this information, we must automatically classify the holders using machine learning techniques. Hence, we obtained $2,715$ patent holders after deduplication, which were categorized into three classes: Company, Science and Technology Institution, and Person. The labeling process was carried out semi-automatically using keyword-based filtering, followed by a manual review of each instance. In total, the training dataset comprises $1{,}430$ companies, $1{,}166$ individuals, and 119 institutions.

Furthermore, we train a model to classify all patent holders. For this purpose, we analysed three algorithms: Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Network (DNN). We also include a fastText [Grave et al. 2018] embedding model because of its bilingual capability and effectiveness in handling non-existent or hybrid words, which is beneficial for classifying patent holder names. The Macto F1 of these models was 0.97 for SVM, 0.96 for DNN, and 0.94 for RF. Since DNN and SVM exhibited statistically similar results, we opted for the DNN-based approach since the validation methods employed did not show signs of overfitting.

Upon applying the machine learning model, we classified over $67{,}000$ holders according to the categories previously outlined. The result includes $35{,}374$ for companies, $29{,}413$ for individuals, and $3{,}082$ for science and technology institutions. These categories will later be used in Section 5 to identify patterns in technological areas across the different types of relationships.

DIP-BR is publicly available in Zenodo, an open research data repository committed to the principles of open science [da Cruz et al. 2025]. By making the dataset available, we encourage transparency, collaboration, and further data exploration by the research community. Finally, regarding its format, our main dataset is available in **JSON files** and the clustering IDs are provided separatly in a **CSV file**. These files formats simplifiy data processing in popular programming languages such as Python and R, enabling the execution of complex analyses and visualizations.

Network Modeling. We model the relationship as a collaboration network to better understand and analyze the relationships between patent holders. To accomplish this, we selected the co-owned patents and converted the classified and deduplicated data into a format compatible with the Gephi network analysis software. This collaboration network is modeled as an undirected graph $G = (V, E)$, where the nodes (V) are patent holders (ICTs, individuals, or companies) and the edges (E) indicate co-ownership of at least one patent between two holders.

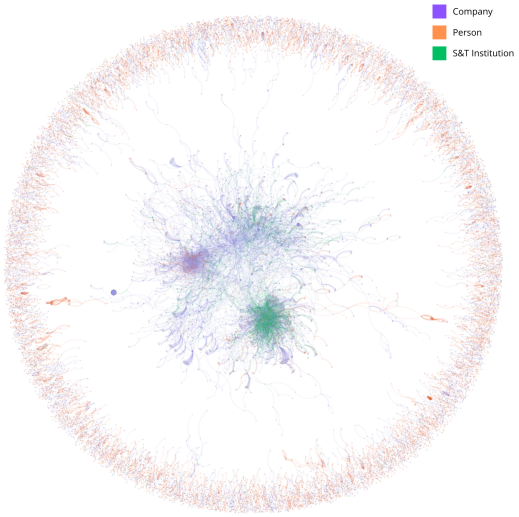

Figure 4. Collaborative patent network overview.

Table 5. Main statistics from the patent network.

Metric	Result
Nodes	18,634
Edges	23,118
Density	0
Average Degree	2.503
Average Clustering Coef.	0.757
Average Length	5,783
No. of Connected Components	4,972
Diameter	18

In this network, the node degree reflects the number of direct collaborations a given entity has, providing insight into how connected different types of holders are. Table 5 presents key metrics extracted from the patent collaboration network registered with the INPI. Note that, differently from Table 4, when modeling the network, only those holders who have collaborated on at least one patent will be considered. The average degree suggests that most patents have only a small number of co-holders, indicating a preference for partnerships involving few entities. Figure 4 provides an overview of the collaborative patent network. The most connected nodes primarily represent companies and S&T institutions, while researchers tend to occupy positions farther from the network's core. This pattern occurs because researchers typically have fewer direct collaborations compared to organizations.

Result Analysis. We grouped the elements based on similar characteristics to conduct an analysis of the processed data. Hence, we applied the K-Means algorithm clustering. The clusters were formed based on the holder type, the total number of patents filed and granted, and the distribution of these patents across IPC classifications. The number of clusters was determined using the Elbow Method, which initially suggested 11 clusters. However, this configuration resulted in many single-entity clusters. To prevent this sparsity from compromising the generalization of our findings, we opted for a more balanced partition of 8 clusters.

5. Results

To analyze our results, we applied clustering techniques. We examined the preprocessed features of each cluster, including the number of patent applications, the number of granted patents, CIP classification, and the collaboration graph.

In the collaboration network, the connections between holders may help identify the most recurring categories in each type of relationship. Table 6 shows the number of collaborative patents by CIP classification for each relationship type.

Table 6. Number of patents in each international classification class (A to H) for each type of relationship existing in the network.

Relation	A	B	C	D	E	F	G	H
Company x Company	5,177	2,267	6,453	244	390	1,060	1,740	1,293
Company x ICT	1,772	520	1,746	36	46	77	688	258
Company x Person	2,219	882	1,608	58	211	369	774	325
ICT x ICT	1,505	316	1,083	14	13	26	476	114
ICT x Person	196	92	142	4	4	5	35	27
Person x Person	3,771	1,448	1,125	43	483	569	991	414

Edges connecting companies are the most common in the network, with patents from these collaborations primarily concentrated in codes A (Human Necessities) and C (Chemistry and Metallurgy). Partnerships involving science and technology institutions (ICTs) account for the smallest proportion of collaborative patents. Despite their lower numbers, these collaborations may link academia and industry, facilitate knowledge transfer, and develop foundational technologies. The relatively low volume of ICT-related patent collaborations could indicate barriers such as bureaucratic challenges, differing innovation timelines, or limited incentives for academia-industry partnerships.

Considering the clustering analysis, Figure 5 shows each cluster's distribution of filed and granted patents by owner type. In contrast, Figure 6 illustrates the CIP distribution across groups. Then, it is possible to observe that each cluster gathers patents with similar CIP code profiles, reflecting the convergence of technological areas and collaborative strategies.

In Figure 5, we could show that companies falls into different clusters. Group 2, for example, represents companies that have filed multiple patents in the physics and electricity fields. However, companies also appear in other clusters, with varying numbers of patent deposits belonging to different CIPs. Research institutions belong to three groups based on their patent filings/grants: Group 3 includes those with fewer patents, Group 0 comprises institutions with a dozen patents, and Group 6 represents those with over a hundred patent deposits.

Analysing also Figure 6 we can see that group 0 predominates institutions and companies, with a balanced distribution across the CIP codes. However, this group also includes many individuals with sporadic contributions. On the other hand, group 1 highlights a convergence between companies and institutions, with a concentration of CIP codes in categories A, B, and C. The variation in the number of individuals grouped within the clusters is related to the different interactions between companies, science and technology institutions, and individuals. Group 2 includes companies that have produced multiple patents in the fields of physics and electricity, while Group 7 is mainly associated with the Chemistry and metallurgy category (CIP code C).

6. Conclusion

Patent analysis uncovers an understanding of innovation dynamics, identifying technological trends, and supporting decision-making for funding agencies and policymakers. However, the study of relationships through co-ownership remains an underexplored approach, offering new insights into collaboration patterns within the innovation ecosystem. Understanding these interactions can help stakeholders identify bottlenecks in technolog-

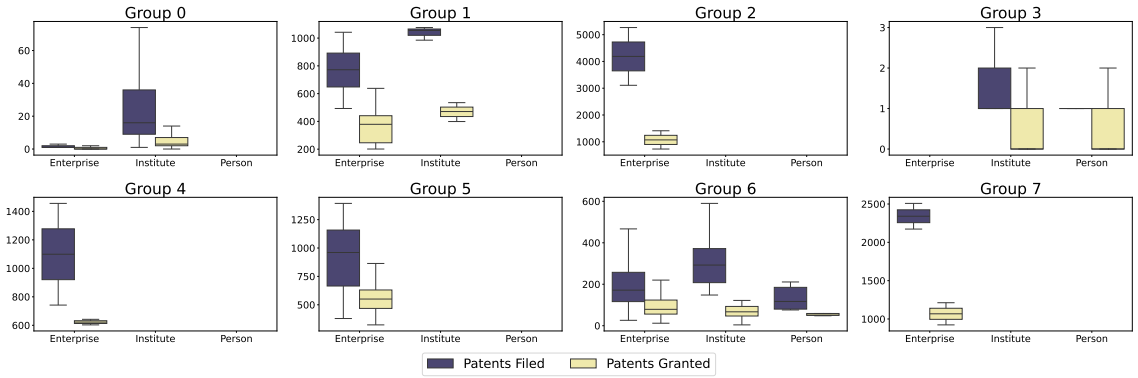

Figure 5. Boxplot Analysis of Patent Collaboration: Identifying central trends, dispersion, and outliers among different clusters.

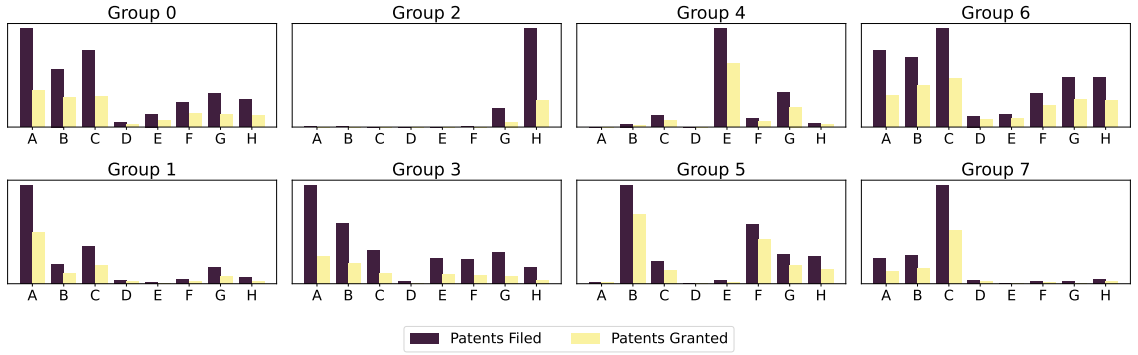

Figure 6. Distribution of patents according to the International classification by groups.

ical production and develop strategies to foster stronger connections between academia, industry, and independent inventors.

In this study, we crawled and enriched data from INPI bulletins, incorporating: classification of patent holder types, clustering analysis, and network modeling. Our dataset includes over 200,000 patents and approximately 70,000 patent holders. The results underscore the value of this dataset for analyzing and understanding the dynamics between research institutions and companies. Our analysis reveals a persistent gap between science and technology institutions and companies, both in individual patent production and collaborative initiatives. This disparity highlights the need for policies that promote stronger academia–industry partnerships to enhance knowledge transfer and drive innovation. Moreover, grouping patent holders using clustering techniques allowed us to uncover patterns of similarity, not only in terms of the number of patents filed or granted but also in the concentration of technological fields. This segmentation sheds light on how different actors engage in innovation and how their collaborative strategies vary across sectors.

Limitations. INPI bulletins have been available in XML format only since 2018, restricting the data collection period and potentially limiting the historical analysis of collaboration trends. Additionally, the deduplication process was not fully automated due to the complexity of distinguishing similar entity names. For instance, due to name similarities,

clustering techniques grouped all Federal Institutes under a single entity.

Future Works. Future improvements include an automated deduplication approach to improve data accuracy and reliability. We also plan to develop a web-based interactive visualization to better explore collaboration networks, enabling researchers, policymakers, and industry to derive actionable insights from patent data more effectively.

7. Acknowledgements

The authors thank FAPEMIG for the support, through the project *Fortalecimento da relação ICT-Empresa: construção de boas práticas de gestão de Propriedade Intelectual e Transferência de Tecnologia* (ACN-00070-21). This work was supported by *Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação* (INCT-TILD-IAR) granted by CNPq (408490/2024-1).

References

Abbas, A., Zhang, L., and Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13.

Asitah, N., Purnomo, A., Young, M. N., Prasetyo, Y. T., Anam, F., Persada, S. F., and Kurniawan, B. K. (2024). Business analytics: A patent landscape retrospective mapping. *Procedia Computer Science*, 234:545–552.

Costa, B. M. G., da Silva Florencio, M. N., and de Oliveira Junior, A. M. (2018). Analysis of technological production in biotechnology in northeast brazil. *World Patent Information*, 52:42–49.

da Cruz, P. V., Lima, I. V. A., Seufitelli, D. B., Dalip, D. H., and de Campos, F. P. V. (2025). DIP-BR: Dataset of Intellectual Property in Brazil. *Zenodo*.

da Silveira, F., Ruppenthal, J. E., Lermen, F. H., Machado, F. M., and Amaral, F. G. (2021). Technologies used in agricultural machinery engines that contribute to the reduction of atmospheric emissions: A patent analysis in brazil. *World Patent Information*, 64:102023.

de Almeida Chaves, D. S., de Melo, G. O., and Corrêa, M. F. P. (2019). A review of recent patents regarding antithrombotic drugs derived from natural products. *Studies in Natural Products Chemistry*, 61:1–47.

Fujino, A. and Stal, E. (2007). Gestão da propriedade intelectual na universidade pública brasileira: diretrizes para licenciamento e comercialização. *Revista de Negócios*, 12(1):104–120.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Kim, G. and Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, 117:228–237.

Matias, A. G. C., Pedreira, D. P., Costa, A. A. N. A., Sanatana, L. T. C., and Santana, V. E. C. (2020). Obtenção de patente e os aspectos do regime de copropriedade. *Revista Brasileira Multidisciplinar*, 23(1):202–213.

Rezende, N. G., Dalip, D. H., Brandão, M. A., and Vasconcelos, M. A. (2023). Elaboraçao de um conjunto de dados sobre o registro de patentes no brasil. In *Dataset Showcase Workshop (DSW)*, pages 99–108. SBC.

Suzgun, M., Melas-Kyriazi, L., Sarkar, S., Kominers, S. D., and Shieber, S. (2023). The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *Advances in neural information processing systems*, 36:57908–57946.

Trappey, C. V., Wu, H.-Y., Taghaboni-Dutta, F., and Trappey, A. J. (2011). Using patent data for technology forecasting: China rfid patent analysis. *Advanced Engineering Informatics*, 25(1):53–64.