

Automated Data Warehouse Architecture to Support Knowledge Management in Patent Databases

Rodrigo Ramos Nogueira, Alexandre Leopoldo Gonçalves, Tobias Felipe Kiefer

Programa de Pós Graduação em Engenharia, Gestão e Mídia do Conhecimento (PPGEGC)

Departamento de Engenharia, Gestão e Mídia do Conhecimento

Universidade Federal de Santa Catarina (UFSC)

rodrigo.nogueira@ifsc.edu.br, a.l.goncalves@ufsc.br, tobias.f.kiefer@gmail.com

Abstract. *This work presents an automated Data Warehouse architecture designed to support the structuring, analysis, and management of knowledge in patent databases. Based on XML files provided by the United States Patent and Trademark Office (USPTO), a Python-based Extract, Transform, Load (ETL) pipeline was developed to preprocess, standardize, and load data into a PostgreSQL database modeled using a star schema. The architecture, executed through Docker containers, integrates a staging layer, analytical dimensions, a fact table, a REST API, and OLAP queries. In the initial load, 189,410 raw records were processed and consolidated into 16,311 unique patents and 651,117 analytical associations. Textual processing identified 25,658 distinct terms through tokenization, while Artificial Intelligence-based semantic enrichment added 45,449 technical terms. The analyses supported the exploration of term frequencies, temporal evolution, technological categories, thematic clusters, semantic co-occurrences, and an exploratory demonstration of normalized temporal forecasting. The results demonstrate that the multidimensional model enables temporal, semantic, technological, geographical, and authorship perspectives to be combined within an integrated structure. Therefore, the proposed solution provides a reproducible and extensible analytical foundation for supporting trend monitoring and future validated forecasting studies, trend monitoring, and decision-making in research, development, and innovation.*

1. Introduction

In a context marked by intense digital transformation, technological innovation has become one of the main drivers of competitiveness among companies, universities, and research institutions. In this scenario, access to reliable and up-to-date technical information assumes a strategic role, particularly for monitoring technological changes and supporting decision-making processes. Among the different sources available, patent databases stand out for providing organized and structured information on inventions, technologies, and discoveries developed across different fields of knowledge. In addition to technical descriptions, these repositories provide relevant metadata, such as authors, filing dates, and technological classifications, which can be explored to identify trends, map emerging areas, and support strategic and predictive analyses [Nogueira and Gonçalves 2025].

The appropriate use of these databases makes it possible to identify emerging trends, map innovation trajectories, and support knowledge management in corporate and academic environments [González, Sakata, and Nogueira 2020]. In developed countries, patent databases are already widely explored for technology forecasting, competitive mapping, and support for research and development. However, access to and effective use of these databases still face significant technical barriers, particularly in environments with limited computational or human resources.

Despite their strategic relevance, extracting informational value from these databases is a nontrivial challenge due to the large volume of data, the complexity of XML formats, and the need for advanced techniques for data organization and analysis. USPTO patent files are provided in weekly XML packages, with a dense hierarchical structure that is often inconsistent across

versions, making the automated extraction process more complex and susceptible to interpretation errors and the loss of relevant data [Krestel et al. 2021]. Approximately 1 GB of data containing thousands of highly detailed technical records is released each week. Much of the difficulty lies in the semi-structured nature of these XML files, which require specialized technical knowledge for data extraction and transformation.

A Data Warehouse (DW) integrates data from multiple sources while maintaining characteristics such as subject orientation, non-volatility, and time variance [Kimball and Ross 2013], making it suitable for storing and analyzing large volumes of historical and heterogeneous data, such as patent records provided by the USPTO. In this context, this work proposes an automated architecture to support knowledge management and strategic decision-making, comprising: (i) a Python-based ETL pipeline to extract, clean, standardize, and transform XML files; (ii) a multidimensional Data Warehouse organized into fact and dimension tables to enable temporal, comparative, and segmented analyses; and (iii) a REST API supporting OLAP queries and access to AI-enriched patent data. Process automation reduces update costs, improves reliability, and enables continuous cycles of data ingestion and analysis.

In the implemented pipeline, Artificial Intelligence and Natural Language Processing techniques were used specifically for the extraction and semantic enrichment of technical terms from patent titles and abstracts. This integration expands the informational value of the structured data and enables the Data Warehouse to support thematic, temporal, and relational analyses. In this regard, the work presents a practical and replicable approach for projects involving large volumes of specialized textual data.

2. Theoretical Background and State of the Art

2.1 Patents as a Source of Technological Knowledge

Patents constitute one of the most structured and accessible public sources of technological knowledge. To obtain protection for an invention, the inventor or responsible organization must provide sufficiently detailed information about its operation, applications, and novel elements. This disclosure is part of the patent-granting process, in which the State grants the holder a temporary right of exclusive exploitation in exchange for the publication of the technical knowledge associated with the invention. In this context, patents serve a dual function: they act as legal instruments for protecting intellectual property while also serving as repositories of scientific and technological information. For this reason, patent documents can be explored in studies aimed at identifying innovations, analyzing technological evolution, and supporting the formulation of organizational strategies [OECD, 2009; WIPO, 2025].

Furthermore, patent documents contain widely standardized documentary and classification elements, following international agreements such as the Patent Cooperation Treaty (PCT), which facilitates their automated processing. Databases such as the United States Patent and Trademark Office (USPTO), Espacenet, maintained by the European Patent Office (EPO), and the Brazilian National Institute of Industrial Property (INPI) provide thousands of records containing metadata such as titles, abstracts, technical fields, inventors, filing dates, and international classifications. Despite this organization, the analytical use of these data still poses significant challenges. The XML structure of patent documents requires specific tools for data reading and extraction, whereas the textual nature of abstracts and descriptions demands Natural

Language Processing (NLP) techniques to identify patterns, relevant concepts, and semantic relationships.

From a knowledge management perspective, patent data constitute an information asset of strategic value. When integrated and organized in a Data Warehouse supporting temporal and thematic analyses, they can inform decisions related to research, development, and innovation. However, limitations involving access, integration, and processing still contribute to the underuse of this potential, particularly in applications requiring large-scale analysis, such as text mining, technological trend identification, and support for the formulation of innovation policies [González, Sakata, and Nogueira 2020].

2.2 Data Warehouse Architecture

A Data Warehouse (DW) is a data structure designed to support decision-making by storing and integrating historical data from different sources and organizing them to facilitate complex analytical queries and the generation of management reports. Unlike Online Transaction Processing (OLTP) systems, which are designed to record and process daily operations, a DW prioritizes the analysis of large volumes of data over time. According to Inmon [2005], a Data Warehouse is characterized as subject-oriented, integrated, non-volatile, and time-variant, properties that structurally distinguish it from operational systems.

Subject orientation indicates that data are organized around topics relevant to the analyzed domain, which, in this work, corresponds to patent data. Integration refers to the consolidation and standardization of data from heterogeneous sources, enabling their unified analysis. Non-volatility, in turn, establishes that, once loaded, data are preserved and do not undergo frequent changes. Time variance enables the maintenance of historical records and their organization by dates and periods, a characteristic that is particularly relevant for identifying and analyzing technological trends over time.

2.3 Dimensional Modeling and Star Schema

Dimensional modeling is one of the most widely used techniques for structuring data in Data Warehouse environments. Unlike normalized models employed in transactional systems, this approach prioritizes simplicity, readability, and query performance, facilitating the analysis of large volumes of data. Its most common representation is the star schema, which consists of a central fact table connected to different dimension tables [Kimball and Ross 2013].

The fact table stores quantitative measures, such as values, counts, and occurrence metrics, which can be analyzed from different perspectives. Dimension tables, in turn, contain descriptive attributes that provide the context required for analysis, such as time, location, authorship, and classification. According to Nogueira [2020], the star schema constitutes a simple and efficient structure for Business Intelligence (BI) environments because it reduces the complexity of table joins and facilitates navigation across dimensions.

In the context of this work, the fact table records the number of word occurrences in patent abstracts, whereas the dimensions represent words, authors, dates, countries, and document metadata. This organization enables Online Analytical Processing (OLAP) operations, such as temporal analyses, the generation of technical term rankings, and the identification of patterns in the data.

2.4 ETL Processes and Textual Noise Reduction

The Extract, Transform, Load (ETL) process constitutes one of the fundamental stages in the construction and maintenance of a Data Warehouse. Its function is to extract raw data from different sources, subject them to transformation procedures that ensure quality and standardization, and subsequently load them into structures designed for analysis.

According to Golfarelli and Rizzi [2009], the ETL process can be organized into three phases: (i) extraction, responsible for collecting data from relational databases, spreadsheets, Application Programming Interfaces (APIs), and semi-structured files such as XML and JSON; (ii) transformation, which comprises noise removal, inconsistency correction, format standardization, attribute derivation, and the application of business rules; and (iii) loading, in which the transformed data are inserted into the analytical structure, often with the support of an intermediate area known as staging. In applications involving textual data, the transformation stage is particularly relevant because it reduces elements that do not contribute to the analysis and prepares the content for subsequent processing stages. In this context, Kimball and Ross [2013] emphasize that ETL-related activities may account for between 60% and 80% of the total effort involved in Data Warehouse projects.

In textual analysis tasks, the transformation stage is essential for preparing data before loading them into the Data Warehouse. This stage includes activities such as cleaning, standardization, stopword removal, noise reduction, and semantic enrichment, which adds context and meaning to the extracted terms. As a result, the data become more suitable for frequency, association, classification, and trend analysis [González, Sakata, and Nogueira 2020].

2.5 Artificial Intelligence Models for Patent Text Analysis

The analysis of patent textual data presents challenges related to the volume of unstructured information, semantic complexity, and the technical vocabulary used in the documents. In this context, Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques have been employed to expand the capacity to process these databases and extract knowledge from them. The main applications include word vector representations, text classification, keyword extraction, and automatic summarization, which enable the identification of relevant concepts, patterns, and information in extensive patent abstracts and descriptions [Lin and Chou 2025].

According to Lin and Chou [2025], Artificial Intelligence techniques applied to patent analysis can support tasks such as automatic summarization, thematic classification, keyword extraction, and technological trend detection. These applications make it possible to condense abstracts and descriptions, assign technical categories, identify terms that represent inventions, and analyze the temporal evolution and co-occurrence of concepts. When integrated into a patent Data Warehouse, these techniques add a qualitative layer to the analysis of structured data, contributing to knowledge management and competitive intelligence.

Among the techniques used for textual analysis, word vector representations, commonly referred to as word embeddings, enable textual tokens to be transformed into continuous numerical vectors that preserve semantic and syntactic relationships among terms. Word2Vec, proposed by Mikolov et al. [2013], learns dense representations from the occurrence of words within local context windows, making it possible to identify similarities and relationships within large textual corpora. Pennington, Socher and Manning [2014] introduced GloVe, which combines local contextual information with global co-occurrence statistics from the corpus. Joulin et al. [2016],

in turn, proposed fastText, which represents words through combinations of character n-grams. This characteristic improves the representation of rare and out-of-vocabulary terms and is particularly relevant in specialized domains, where technical expressions and morphological variations occur frequently, as in patent abstracts and descriptions.

More recently, contextual language models based on the Transformer architecture have expanded the ability to represent textual meaning by considering the context in which each term occurs. BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. [2019], generates context-dependent representations and has achieved significant results in tasks such as text classification, named entity recognition, and semantic similarity. Radford et al. [2018] proposed GPT (Generative Pre-trained Transformer), a generative architecture pretrained on large textual corpora and subsequently adapted to specific tasks, including text generation and automatic summarization. Unlike static embeddings, contextual models can assign different representations to the same word according to its use within a sentence, which is especially important for the ambiguous and domain-specific vocabulary found in patent documents.

In the context of this work, these models support the semantic enrichment stage of the proposed architecture. The terms extracted or generated from patent titles and abstracts are integrated into the multidimensional structure of the Data Warehouse, complementing the structured attributes already stored in its dimensions. This integration makes it possible to explore semantic information together with temporal, technological, geographical, and authorship perspectives, expanding the analytical possibilities beyond isolated textual processing.

2.6 Related Work

In the context of patent data analysis, different studies have explored components associated with the construction of analytical environments, such as ETL processes, relational databases, dimensional modeling, and Data Warehouse architectures. Although not all proposals integrate these elements into a single solution, their results provide important support for the technological and architectural choices adopted in this work.

Silva [2023] analyzes the performance of PostgreSQL, a relational database, and Neo4j, a graph-oriented database, in accessing complex patent citation data. The results indicate an advantage for PostgreSQL in operations involving aggregations and simple traversals, reinforcing the technological choice adopted in this work. Complementarily, Almeida [2011] investigates the feasibility of open-source solutions, such as PostgreSQL, for building low-cost Data Warehouses, demonstrating that, although the system has limitations in scenarios requiring extreme scalability, it provides satisfactory performance in real-world applications with moderate computational resources.

Machado et al. [2019] present DOD-ETL, a distributed and event-driven architecture designed for near-real-time data loading, achieving significant improvements in latency and throughput compared with traditional batch ETL approaches. Zhang [2020], in turn, describes the construction of a Python-based ETL pipeline integrated with a PostgreSQL database modeled using a star schema. The study details strategies for extracting and transforming local data into a relational environment, presenting an approach similar to that adopted in this research.

Kong [2015] proposes an architecture for publishing open data based on ETL processes and Data Warehousing, applicable to government systems. The results demonstrate the feasibility of using structured data to support public management and highlight standardization challenges similar to those found in patent databases. Collectively, these studies reinforce the architectural

choices of this project and indicate gaps related to the automated processing of large volumes of patent data in XML format and the integration of these data with textual analysis techniques.

3. Materials and Methods

This work is characterized as applied research of a technological and exploratory nature [Junior et al. 2014], aimed at developing and validating an automated Data Warehouse architecture for collecting, structuring, and analyzing data extracted from patent documents. The methodological process was organized into four sequential stages: (i) multidimensional data modeling; (ii) implementation of an automated Extract, Transform, Load (ETL) pipeline; (iii) development of an Application Programming Interface (API) for data access; and (iv) exploratory analysis of the loaded data through Online Analytical Processing (OLAP) queries.

The Data Warehouse model was defined based on an analysis of the structure of the XML files provided by the United States Patent and Trademark Office (USPTO). These files contain complex and hierarchically organized data, including technical abstracts, classifications, dates, inventors, and geographical information. To integrate and organize these data, a star schema was adopted following the dimensional modeling recommendations proposed by Kimball and Ross [2013].

Data extraction, transformation, and loading were performed through an ETL pipeline developed in Python. During the extraction stage, the pipeline automates the reading of XML files provided by the USPTO and retrieves the fields relevant to the analysis. The transformation stage includes the removal of repetitive metadata, date standardization, name normalization, tokenization, and text cleaning. English stopwords are also removed using the Natural Language Toolkit (NLTK) library, converting the technical content into structures suitable for analytical processing. Finally, the transformed data are initially stored in an intermediate staging area and subsequently distributed across the fact and dimension tables, following strategies designed to handle large volumes of semi-structured data [González, Sakata, and Nogueira 2020].

The proposed architecture includes a RESTful API developed using the FastAPI framework, which provides external access to the consolidated data through filters related to countries, authors, periods, and keywords. To explore the stored data, analytical SQL queries supporting Online Analytical Processing (OLAP) operations were developed, enabling the analysis of temporal patterns, associations among terms, and the geographical distribution of patent records. An AI-based semantic enrichment module was also integrated specifically to extract technical terms from patent titles and abstracts. The generated terms were stored in the **dim_words** dimension and associated with their source documents through the **fact_patents** table. The entire execution environment was containerized using Docker, ensuring reproducibility, dependency isolation, and solution portability. The following sections describe each component of the architecture in greater detail.

3.1 Project Architecture Overview

The proposed architecture was developed to automate the collection, transformation, structuring, and analysis of data extracted from XML files provided by the United States Patent and Trademark Office (USPTO). The data flow begins with the reading of patent application records, which

undergo preprocessing, transformation, and semantic enrichment stages. The resulting data are then loaded into two layers: an intermediate staging area and the Data Warehouse, composed of fact and dimension tables. The entire environment was containerized using Docker, ensuring dependency isolation, experiment reproducibility, and solution portability.

The Extract, Transform, Load (ETL) pipeline was developed entirely in Python and organized into modules corresponding to the three stages of the process, as shown in Figure 1. During the extraction stage, the USPTO XML files are initially processed by a script named `fix_xml`, which removes duplicate declarations and encloses the records within a single root element, enabling their parsing. The `ElementTree` library is then used to navigate the hierarchical structure of the documents and extract information such as invention title, document number, country, filing date, inventors, abstract, and technical description. Because the file structure may vary across versions and contain missing or malformed fields, error-handling mechanisms were implemented to prevent interruptions in the ingestion process and reduce the loss of relevant data.

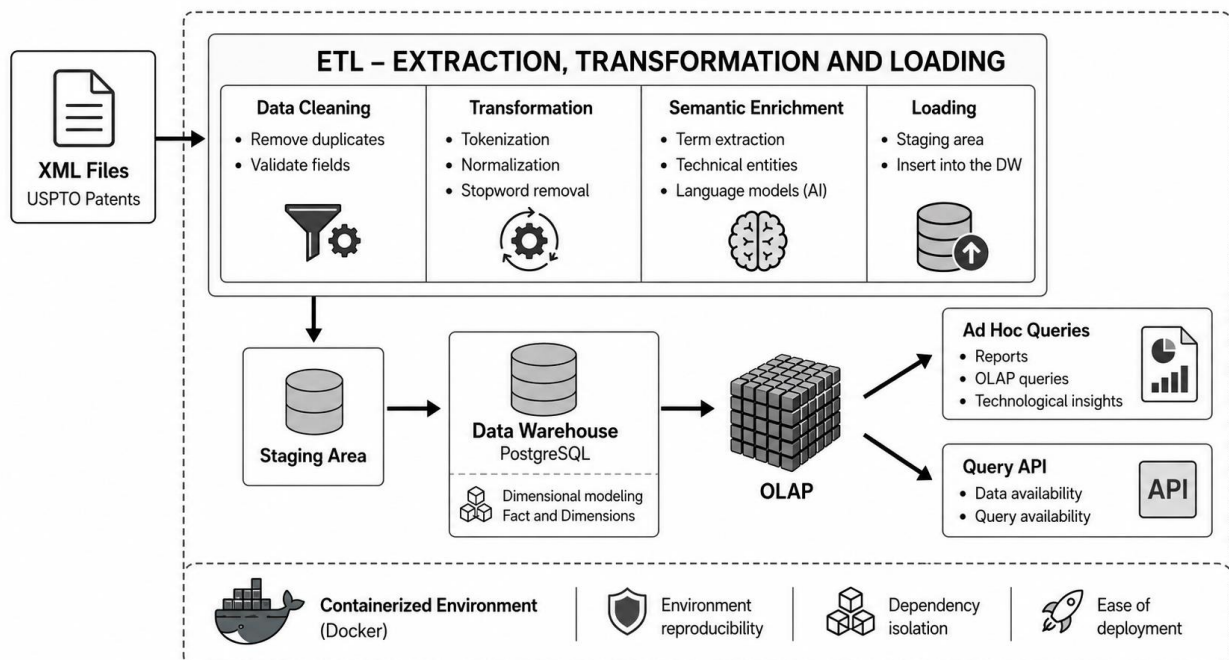


Figure 1 - Project Architecture. Source: Prepared by the authors [2026].

During the transformation stage, the extracted data undergo normalization and textual preprocessing procedures. Inventor names are standardized, dates are converted into temporal structures using the ISO format, and abstracts are segmented into words. The content is then converted to lowercase, punctuation is removed, and English stopwords provided by the Natural Language Toolkit (NLTK) library are filtered. Based on this processing, a set of unique words is generated for each abstract, reducing the presence of terms with low semantic value and preparing the data for subsequent analyses. As a future extension, lemmatization or stemming techniques may be incorporated to improve the normalization and semantic quality of the stored terms.

During the loading stage, the transformed data are initially inserted into the intermediate ``staging_patents`` table. From this structure, the records are distributed across the dimension tables and the Data Warehouse fact table. Loading is performed incrementally, with prior verification of whether records already exist in the dimensions, preventing duplicates and ensuring referential integrity during execution. The ``psycopg2`` package is used to connect to PostgreSQL and execute SQL statements. The entire process runs automatically in Docker containers, enabling new files to

be processed with minimal manual intervention, greater scalability, and reduced operational rework.

3.2 Data Warehouse Modeling

The Data Warehouse was modeled using a dimensional star schema approach to facilitate analytical queries and the generation of temporal and thematic summaries. Figure 2 presents the adopted dimensional model, highlighting the central fact table, its dimensions, and their corresponding relationships.

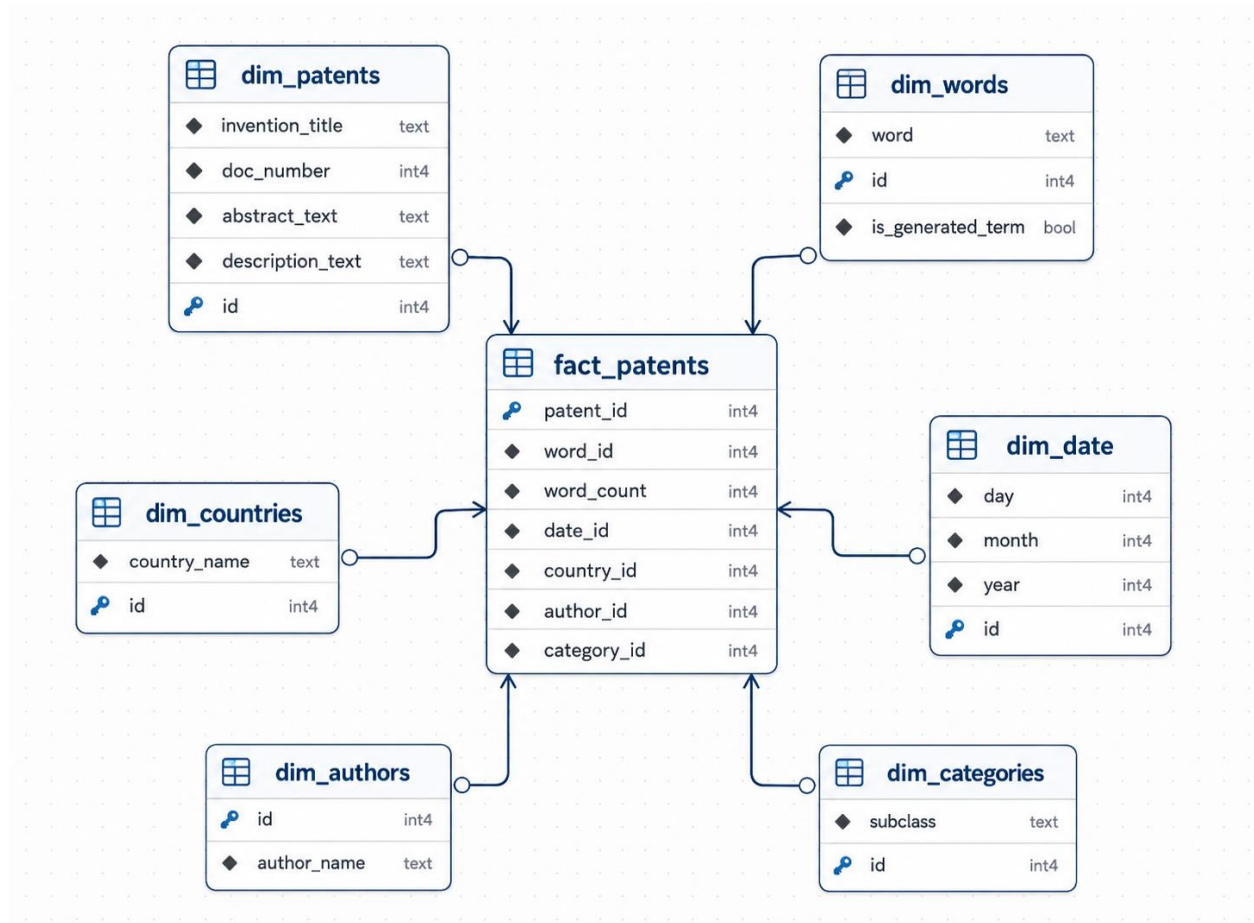


Figure 2 - Dimensional model adopted in the patent Data Warehouse. Source: Prepared by the authors [2026].

The fact_patents fact table stores the occurrences of words identified in patent abstracts, using the word_count attribute as the quantitative measure of the analysis. This table is connected, through foreign keys, to six dimensions: patents (dim_patents), words (dim_words), dates (dim_date), countries (dim_countries), authors (dim_authors), and technological categories (dim_categories). The dim_patents dimension stores the invention title, document number, abstract, and technical description. The dim_words dimension records the extracted terms and indicates whether each word was obtained directly from the text or generated through semantic enrichment processes. The dim_date dimension organizes the records by day, month, and year, while dim_countries and dim_authors store, respectively, the country associated with the document and the names of the inventors. The dim_categories dimension represents the technological subclasses associated with the patents.

This structure supports different analytical questions, such as identifying the most frequent terms by technological category or period, determining which countries concentrate the largest number of records, analyzing inventor activity, and observing the temporal evolution of specific concepts.

3.3 Environment and Automation with Docker

To ensure reproducibility and facilitate the deployment of the architecture across different computing environments, Docker and Docker Compose were used. The solution consists of three integrated containers: (i) `postgres_patents`, responsible for running the PostgreSQL database and automatically creating the initial schema through the `init_schema.sql` script; (ii) `etl_patents`, which starts the ETL pipeline after the database becomes available and performs the processing and loading of the XML files; and (iii) `api_patents`, which provides access to the processed data through a REST API running on port 8000.

The volumes defined in the `docker-compose.yml` file ensure data persistence across container restarts and allow the source code to remain decoupled from the execution environment. This organization reduces the need for manual configuration, facilitates version control, and makes the deployment process more consistent. Furthermore, the architecture supports the portability and future expansion of the solution, enabling its migration to cloud environments and managed PostgreSQL services, such as Supabase and Amazon Relational Database Service (Amazon RDS).

3.4 REST API for Data Querying

The REST API was developed using the FastAPI framework and runs independently in a Docker container on port 8000. Its structure organizes routes by domain, with specific modules for words, authors, and countries, while SQL queries and the PostgreSQL connection are centralized to facilitate code maintenance and reuse.

The endpoints support queries for the most frequent words, their distribution by year, country, and author, annual rankings, and associations among terms based on co-occurrences in patent abstracts. The API also provides access to patents linked to specific authors and countries. The routes perform OLAP operations on the dimensional model using aggregations, groupings, sorting, and window functions. Interactive documentation is generated automatically through Swagger UI, while the co-occurrence endpoints support the identification of emerging topics and technological trends.

3.5 AI-Based Semantic Enrichment and Dual-Model Evaluation Protocol

The semantic enrichment stage was implemented using the OpenAI API and the gpt-4o-mini language model. Although the proposed architecture supports the integration of different AI providers, all AI-generated terms reported in this study were produced using this specific model. The model received the textual content of each patent, including its title and abstract, and was instructed to identify technical concepts, techniques, and technologies relevant to patent analysis. Authentication credentials, including the API key, were stored outside the source code and are not reported for security reasons.

The prompt was designed to prioritize terminology compatible with the International Patent Classification (IPC) and Cooperative Patent Classification (CPC) domains. It required singular forms, the aggregation of highly similar expressions under a representative term, the exclusion of general English words and nonspecific concepts, and the prioritization of multiword

technical expressions. The model was limited to a maximum of seven terms per patent and was required to return only a Python list, without explanations or additional formatting. The complete instruction used during the enrichment process is reproduced in Box 1.

Box 1 – Prompt used for AI-based semantic enrichment. Source: Prepared by the authors [2026].

Role and objective: Act as a patent analysis expert with deep knowledge of the IPC (International Patent Classification) and CPC (Cooperative Patent Classification) taxonomies. Using the provided text document, extract the most relevant technical terms, such as key concepts, techniques, and technologies, that can be characterized as entities within the context of patent analysis. Rules: Only singular forms of terms are included. Highly similar terms, including synonyms, plurals, or minor variations, must be aggregated under a single representative term. Common English words, including general verbs, adjectives, prepositions, and non-domain-specific nouns, must be excluded. Simple words that do not represent a specific concept, technique, or technology must also be excluded, unless they are acronyms or recognized technical terms. Terms aligned with IPC and CPC classifications must be prioritized. Up to seven terms should be extracted, mainly expressions containing at least two words. Output: Return a Python list containing only the extracted and processed terms, without additional text or formatting.

The origin of each term was preserved in the `dim_words` dimension through the Boolean attribute `is_generated_term`, making it possible to distinguish expressions produced by the language model from terms obtained through conventional tokenization. This attribute also enabled the subsequent evaluation to be restricted specifically to the vocabulary generated during the semantic enrichment stage.

The quality-assessment procedure followed an independent dual-model LLM-as-a-Judge protocol. A deterministic sample of 400 patent–term associations was initially selected using hash-based ordering, corresponding to 395 distinct patents. To preserve the complete semantic context of each selected patent, all AI-generated terms associated with these documents were retrieved and grouped with their respective titles and abstracts. This procedure produced 2,169 patent–term associations for evaluation.

The same evaluation instructions and textual evidence were independently submitted to GPT-5 and Claude Sonnet 5. The judging models did not have access to one another's responses. Each extracted term was classified as valid, partially valid, or hallucinated. A term was considered valid when it represented a technically meaningful concept directly supported by the patent title or abstract; partially valid when it was related to the document but excessively broad, ambiguous, or imprecise; and hallucinated when it was not supported by the supplied textual evidence. For each classification, the models were also requested to return a confidence score, a supporting excerpt, and a brief justification.

Evaluation consistency was examined using raw agreement and Cohen's kappa coefficient. Associations receiving divergent classifications were reserved for subsequent human adjudication. Because the initial assessment was conducted using language models rather than domain experts, it was treated as a reproducible dual-model quality assessment.

3.6 Computational Environment, Reproducibility, and Performance Measurement

The ETL and database services were organized in a containerized architecture using Docker Engine 29.8.1. The container configuration specifies the application dependencies,

database connection parameters, and execution environment required to reproduce the workflow. Performance measurements were conducted directly on the host operating system using the same data structures, processing rules, and analytical queries adopted in the study.

The computational environment consisted of Ubuntu 22.04.5 LTS with a 64-bit Linux 6.8 kernel, an Intel Core i5-10210U processor with four physical cores and eight threads, 8 GB of RAM, and a 512 GB NVMe solid-state drive. The software environment included Python 3.10.12, pandas 2.3.3, NumPy 2.2.6, scikit-learn 1.7.2, psycopg2-binary 2.9.12, the OpenAI Python SDK 2.41.1, and PostgreSQL 14.24.

The PostgreSQL instance was configured with 128 MB of shared buffers, 4 MB of working memory, 64 MB of maintenance working memory, an effective cache size of 4 GB, and a maximum of 100 concurrent connections. The resulting database occupied approximately 2.96 GB. The staging table accounted for approximately 2.63 GB of total storage, while the patent dimension and fact table occupied approximately 215 MB and 50 MB, respectively, including auxiliary storage and indexes.

Performance evaluation covered the AI-based semantic-enrichment stage and six representative analytical database queries. For semantic enrichment, the collected metrics included processing time, input and output token consumption, response validity, and estimated API cost. Mean values obtained from the benchmark were used to project the processing requirements for the complete analytical dataset of 16,311 patents. The projection assumed sequential execution without parallel requests or retries.

The database benchmark included queries for patent counting, annual document distribution, normalized temporal prevalence, annual emergence of AI-generated terms, term-frequency analysis, and IPC-subclass aggregation. Each query was executed 20 times after two warm-up executions. Client-side execution time was measured in Python and summarized using the median and 95th percentile. PostgreSQL execution plans were also examined using EXPLAIN ANALYZE with buffer statistics.

4. Results and Discussion

4.1 ETL Pipeline Execution and Data Quality

The ETL pipeline was executed in a local environment using a Docker-based containerized architecture and XML files provided by the USPTO. This environment standardized script execution, isolated dependencies, and improved the reproducibility of the extraction, transformation, and loading stages. In the proposed workflow, the `staging_patents` table was used as the initial ingestion layer, receiving the records extracted from the XML files before their transformation and distribution across the dimensional model.

At the end of the ingestion stage, the staging layer stored 189,410 raw records, while the `dim_patents` dimension consolidated 16,311 unique patents, which constituted the dataset used to validate the dimensional model and perform the subsequent analyses. This difference results from the granularity of each structure: `staging_patents` preserves the records extracted from the XML files, whereas the dimension stores each document only once based on its identification number. The `fact_patents` fact table recorded 651,117 associations among patents, terms, and dimensions, covering 16,309 patents with associated terms. The difference of two patents indicates that these documents contained no valid terms after tokenization and filtering.

The dimensional structure organized documents, terms, authors, countries, dates, and technological categories. A total of 71,107 distinct terms, 15,727 authors, and 1,497 dates were identified. The evaluation of the staging layer also indicated 100% completeness for document number, title, country, author, filing date, abstract, description, technological section, and subclass, with no missing values in the analyzed attributes.

During transformation, the abstracts were tokenized, filtered for stopwords, and associated with the `dim_words` dimension, while the fact table preserved term frequencies by patent. On average, each document contained 38.34 distinct terms and 75.02 textual occurrences, ranging from 1 to 579 terms and from 1 to 3,772 occurrences per patent. Tokenization produced 25,658 distinct terms and 1,095,028 occurrences. Artificial Intelligence-based semantic enrichment added 45,449 distinct technical terms and 128,464 occurrences, expanding the conceptual diversity available for analysis.

Based on the measured processing rate, the sequential semantic enrichment of all 16,311 patents was estimated to require approximately 7.97 hours, corresponding to an average throughput of approximately 34 patents per minute. The complete dataset was projected to consume approximately 5.03 million input tokens and 0.58 million output tokens, totaling 5.60 million tokens. The projected processing time ranged from 7.37 to 8.71 hours. No request failures or malformed outputs were observed during the controlled benchmark. These results represent linear estimates for sequential processing and may vary according to network conditions, API availability, rate limits, retry policies, and the use of parallel requests.

The PostgreSQL database occupied approximately 2.96 GB. The staging table accounted for 2.63 GB of total storage, while the patent dimension and fact table occupied approximately 215 MB and 50 MB, respectively, including auxiliary storage and indexes. The simple patent-count query presented a median client-side execution time of 7.04 ms and a 95th-percentile time of 9.55 ms. For the five analytical queries, median execution times ranged from 697.67 to 927.08 ms, while 95th-percentile times ranged from 746.17 to 1,122.95 ms. The normalized temporal analysis of the term *electronic device* was the most computationally demanding query, with a median of 927.08 ms and a 95th-percentile time of 1,122.95 ms. Thus, all evaluated analytical queries presented median response times below one second in the benchmark environment.

4.2 Quality Assessment of AI-Generated Terms

The independent quality assessment conducted by GPT-5 and Claude Sonnet 5 covered the complete evaluation set of 2,169 patent–term associations linked to 395 distinct patents. GPT-5 classified 2,027 associations as valid (93.45%) and 142 as partially valid (6.55%). Claude Sonnet 5 classified 2,164 associations as valid (99.77%) and five as partially valid (0.23%). Neither judging model classified an association as hallucinated.

The two models produced the same classification for 2,028 associations and different classifications for 141, resulting in a raw agreement of 93.50%. Among the concordant cases, 2,025 associations were classified as valid by both models and three were classified as partially valid by both. Among the divergent cases, GPT-5 classified 139 associations as partially valid that Claude Sonnet 5 classified as valid. In the remaining two cases, GPT-5 assigned the valid label and Claude Sonnet 5 assigned the partially valid label.

Despite the high raw agreement, Cohen's kappa was 0.037. This result was affected by the pronounced prevalence of the valid category, which increased the agreement expected by chance and consequently reduced the coefficient. The low kappa value therefore does not directly

contradict the observed agreement of 93.50%, but indicates that the agreement was concentrated almost entirely in a single classification category.

The distribution of the divergent judgments also indicates that GPT-5 adopted a more restrictive criterion than Claude Sonnet 5 when distinguishing valid from partially valid terms. Of the 141 disagreements, 139 resulted from GPT-5 assigning the partially valid label to an association considered valid by Claude Sonnet 5. These associations were reserved for human adjudication so that a final classification could be established without automatically privileging the judgment of either model.

Overall, the dual-model assessment provides evidence that most of the evaluated terms were recognized by both judging models as technically meaningful and supported by the corresponding patent title or abstract. However, these findings should not be interpreted as a direct estimate of extraction precision or as a human-validated hallucination rate. The evaluation was performed on a sample of the AI-generated vocabulary, the classification distribution was strongly imbalanced, and the initial judgments were produced by language models. Final human adjudication of the 141 divergent cases is therefore required before reporting a consolidated human-validated classification.

4.3 Exploratory and Descriptive Analyses of the Enriched Patent Database

Based on the loaded dimensional model, analytical queries were implemented to explore different dimensions of the patent corpus. The most frequent terms query, shown in Figure 3, aggregates records from the fact_patents table by the terms generated during semantic enrichment and sorts them by the sum of their occurrences. The 20 most recurrent terms identified in the latest execution highlight the main technological focuses present in the analyzed dataset.

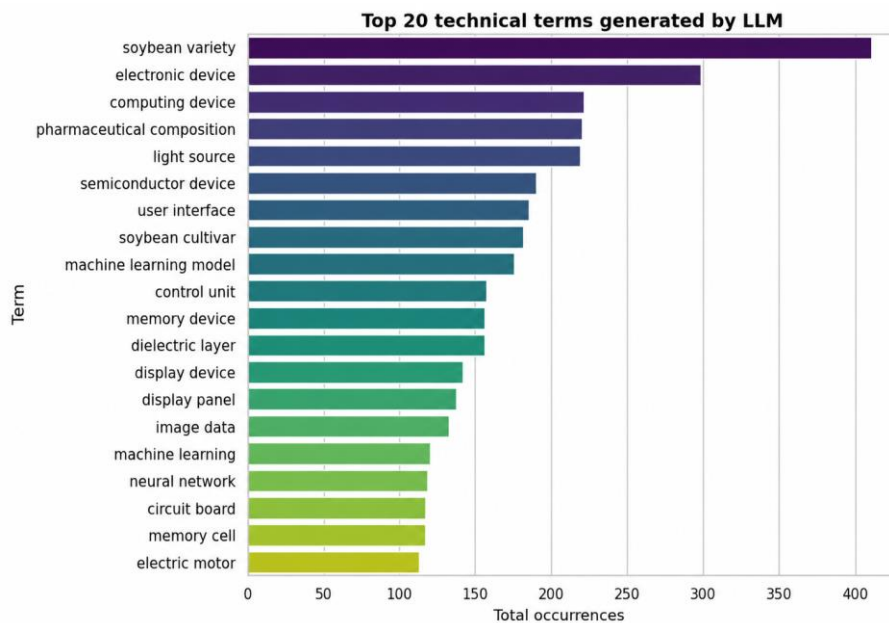


Figure 3 - Top 20 terms generated by semantic enrichment. Source: Prepared by the authors [2026].

Figure 4 presents the normalized annual prevalence of the term electronic device. For each year, the number of distinct patents containing the term was divided by the total number of documents available in that year and expressed per 1,000 documents. Each patent was counted only once, regardless of how many times the term occurred within its textual content. To reduce the instability caused by years with very limited document coverage, the continuous temporal line was restricted to years containing at least 100 patents.

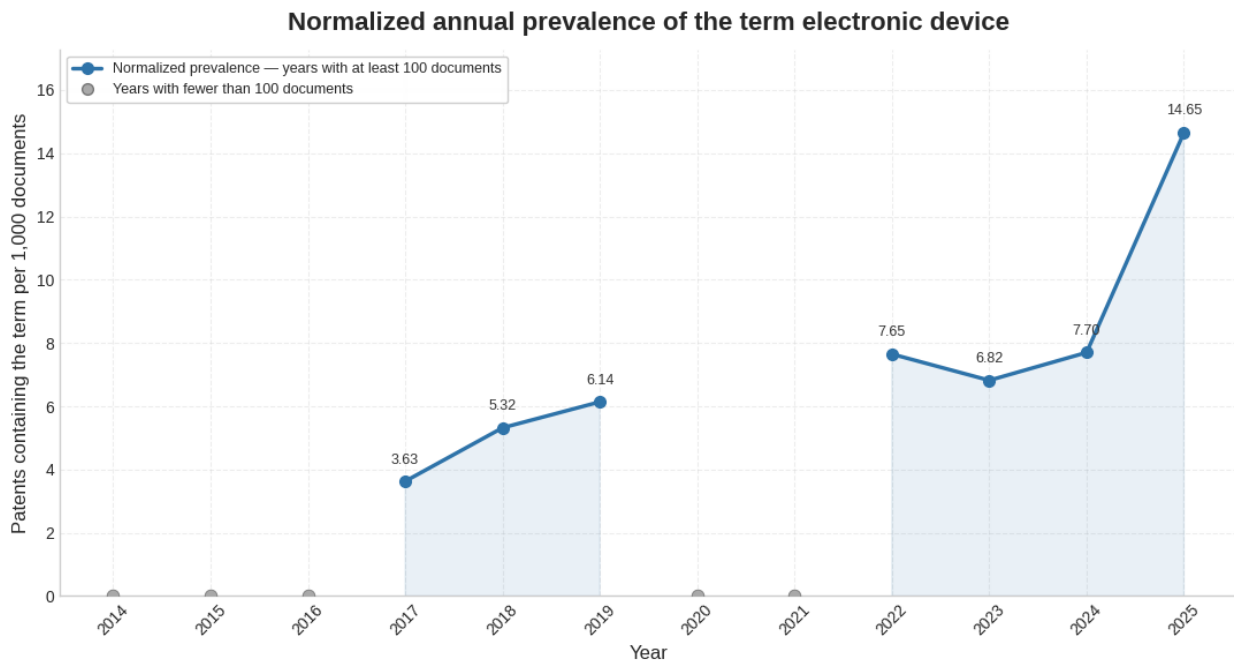


Figure 4 – Normalized annual prevalence of the term electronic device, expressed as the number of patents containing the term per 1,000 documents available in each year. The blue segments represent years with at least 100 documents, while gray markers identify years with more limited coverage. The discontinuity between the two temporal windows reflects the coverage of the analyzed corpus. Source: Prepared by the authors [2026].

The figure presents two interpretable temporal windows, 2017–2019 and 2022–2025, reflecting the limited document coverage in the remaining years rather than an interruption in technological activity. The normalized prevalence increased from 3.63 patents per 1,000 documents in 2017 to 6.14 in 2019 and, after moderate variation between 2022 and 2024, reached its highest value of 14.65 in 2025. These results indicate an increased proportional presence of the term in the available corpus but should be interpreted as descriptive of the analyzed dataset because of its incomplete temporal coverage.

Figure 5 presents the normalized annual emergence of AI-generated technical terms. For each year, the number of terms identified for the first time in the corpus was divided by the number of documents available in that year and expressed per 1,000 documents. Among the years with at least 100 documents, the highest normalized rate occurred in 2017, followed by a gradual reduction through 2019. The rate showed moderate variation between 2022 and 2024 and reached its lowest value among these years in 2025. These results differ from the absolute counts, which were strongly influenced by annual document volume. Although normalization reduces the direct effect of differences in corpus size, it does not eliminate the uncertainty caused by incomplete monthly coverage. The rates for 2014, 2015, 2016, 2020, and 2021 are shown for completeness but should be interpreted with caution because they were calculated from fewer than 100 documents and are therefore highly sensitive to individual records.

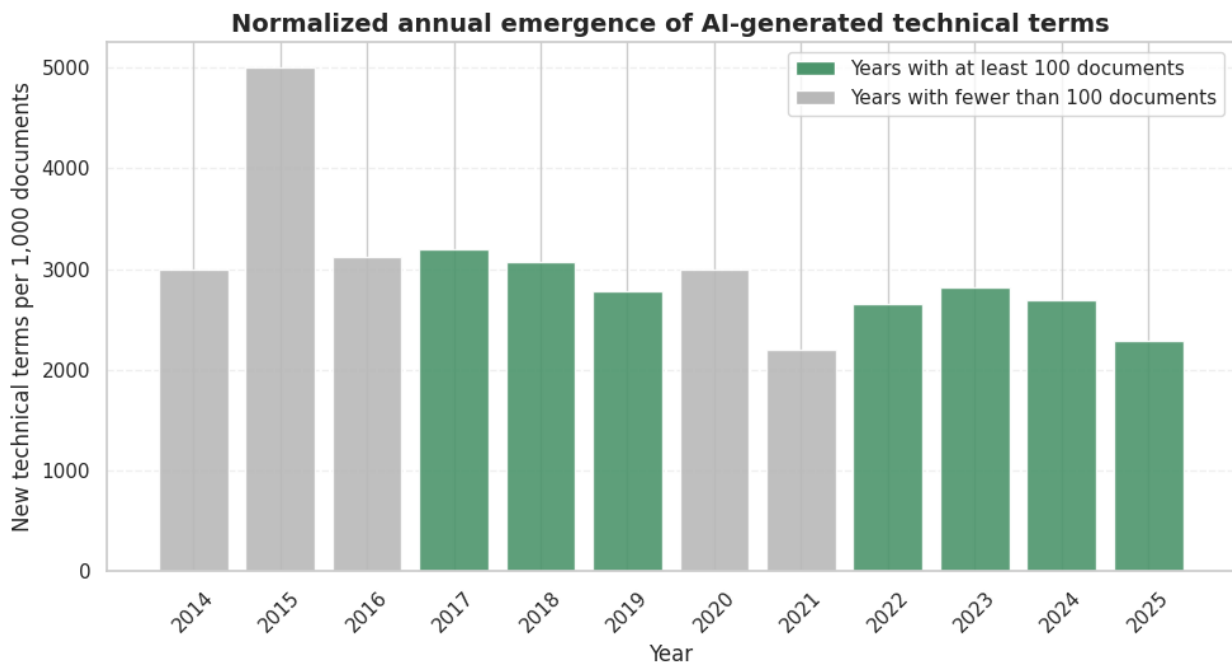


Figure 5 – Normalized annual emergence of AI-generated technical terms, expressed as the number of terms first observed per 1,000 documents available in each year. Gray bars identify years with fewer than 100 documents. Source: Prepared by the authors [2026].

The number of clusters was reassessed using the elbow method, the average silhouette coefficient, and the stability of the resulting partitions. Candidate solutions ranging from $k=2$ to $k=10$ were evaluated. The elbow method indicated an inflection point at $k=4$; however, the highest silhouette coefficient was obtained for $k=3$ (0.0180), compared with 0.0163 for $k=4$ and 0.0164 for the previously adopted $k=5$. Given this divergence, clustering stability was additionally examined through repeated K-Means executions with ten different random seeds and evaluated using the Adjusted Rand Index (ARI). The three-cluster solution achieved a mean ARI of 0.9964, a minimum ARI of 0.9918, and a standard deviation of 0.0016, indicating highly consistent assignments across different initializations. In comparison, the four-cluster solution presented a mean ARI of 0.9048 and a minimum value of 0.5965, whereas the five-cluster solution achieved a mean ARI of 0.9675 and a minimum value of 0.9241. Therefore, although the elbow criterion suggested $k=4$, the combined evidence from relative separation and partition stability supported the selection of $k=3$ as the most reproducible configuration for the revised clustering analysis.

Figure 6 presents the distribution of 16,258 patents across the three thematic clusters identified in the revised analysis. Of the 16,311 unique patents available in the dimensional database, 53 records (0.32%) were excluded because they did not contain any of the 500 terms selected to construct the TF-IDF representation. Cluster 1 comprises 2,614 patents (16.1%) and is predominantly associated with pharmaceutical, biotechnological, and chemical innovations, as indicated by expressions such as pharmaceutical composition, pharmaceutically acceptable salt, tissue culture, nucleic acid, host cell, and amino acid sequence. Cluster 2 is the largest group, comprising 7,097 patents (43.7%), and concentrates technologies related to artificial intelligence, computing, and digital communication, including machine learning model, artificial intelligence, neural network, user interface, wireless communication, and information processing apparatus. Cluster 3 contains 6,547 patents (40.3%) and is characterized by semiconductor technologies, electronic components, and engineered devices, represented by expressions such as semiconductor

device, dielectric layer, semiconductor substrate, circuit board, display panel, and battery module. The thematic characterization was based on compound terms occurring in at least three patents and ranked according to their coverage and concentration within each cluster and their inverse document frequency. Because the silhouette coefficients remained low, these groups should be interpreted as predominant technological profiles rather than as completely isolated thematic categories.

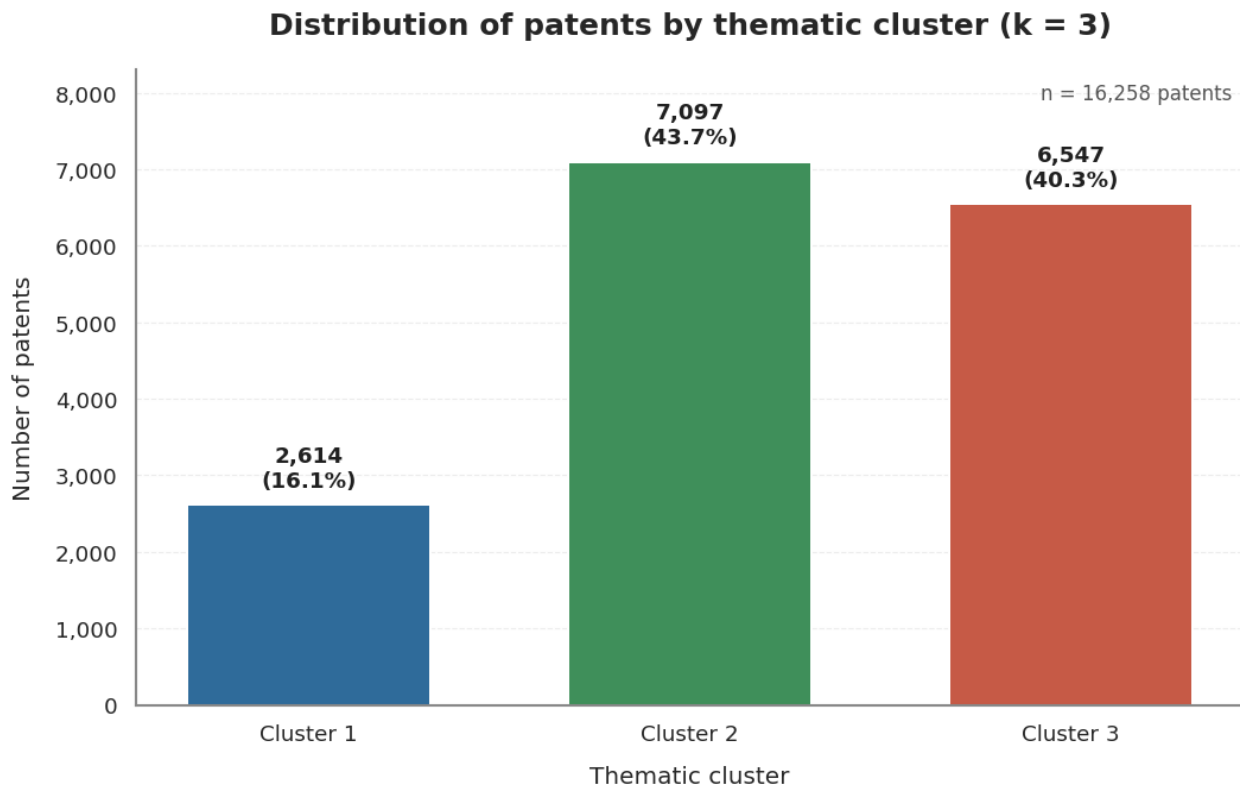


Figure 6 – Distribution of patents across the three thematic clusters identified by K-Means. Source: Prepared by the authors [2026].

Principal Component Analysis (PCA) was applied to the same TF-IDF matrix used in the clustering procedure to provide a two-dimensional exploratory representation of the resulting groups. As shown in Figure 7 (a), the first principal component explains 1.97% of the total variance, while the second explains 1.85%, resulting in a combined explained variance of 3.82%. Therefore, although the projection reveals differences in the spatial concentration of the three clusters, it preserves only a limited portion of the structure contained in the original 500-dimensional feature space. The apparent separation observed in the two-dimensional representation should consequently be interpreted as exploratory rather than as evidence of strongly isolated groups.

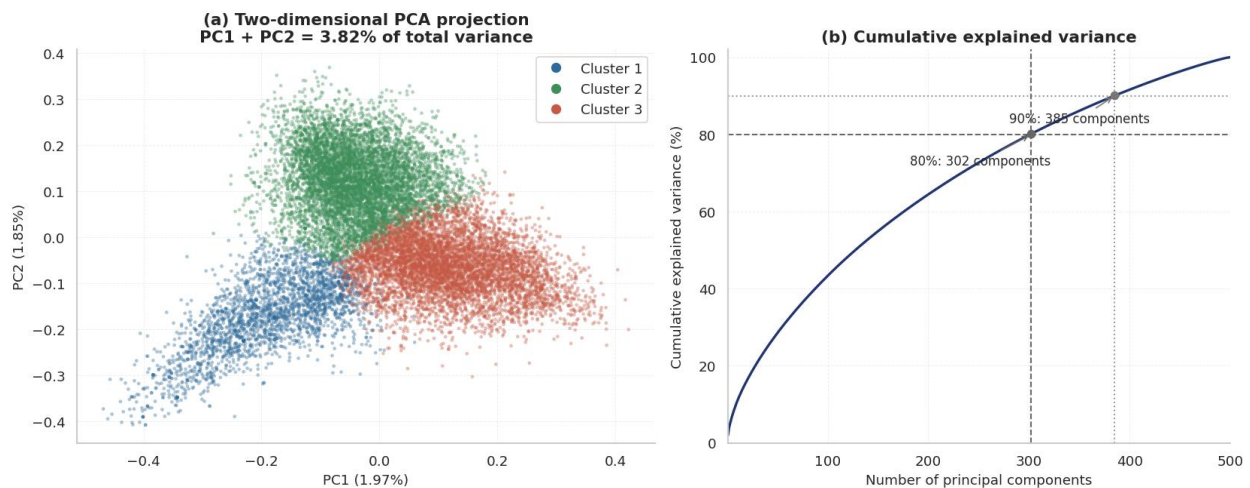


Figure 7 – Two-dimensional PCA projection of the three thematic clusters and cumulative explained variance. The first two principal components explain 3.82% of the total variance. Source: Prepared by the authors [2026].

Figure 7 (b) further demonstrates the high dimensionality of the patent corpus. A total of 302 principal components are required to explain 80% of the variance, while 385 components are necessary to reach 90%. These results indicate that the technological vocabulary represented by the TF-IDF matrix is distributed across a large number of dimensions and cannot be adequately summarized by only two principal components. The PCA projection was therefore retained exclusively as a visual aid, whereas the selection of the three-cluster solution was supported by the combined results of the silhouette coefficient and clustering stability analysis.

4.4 Semantic Co-occurrence and IPC Subclass Analysis

Figure 8 presents the main co-occurrences among technical terms identified within the same patents. The most frequent association occurs between pharmaceutical composition and pharmaceutically acceptable salt, with 22 occurrences. Other notable relationships involve this term and active ingredient (13), nucleic acid (12), and pharmaceutically acceptable carrier (11), indicating a group related to pharmaceutical compositions and formulations.

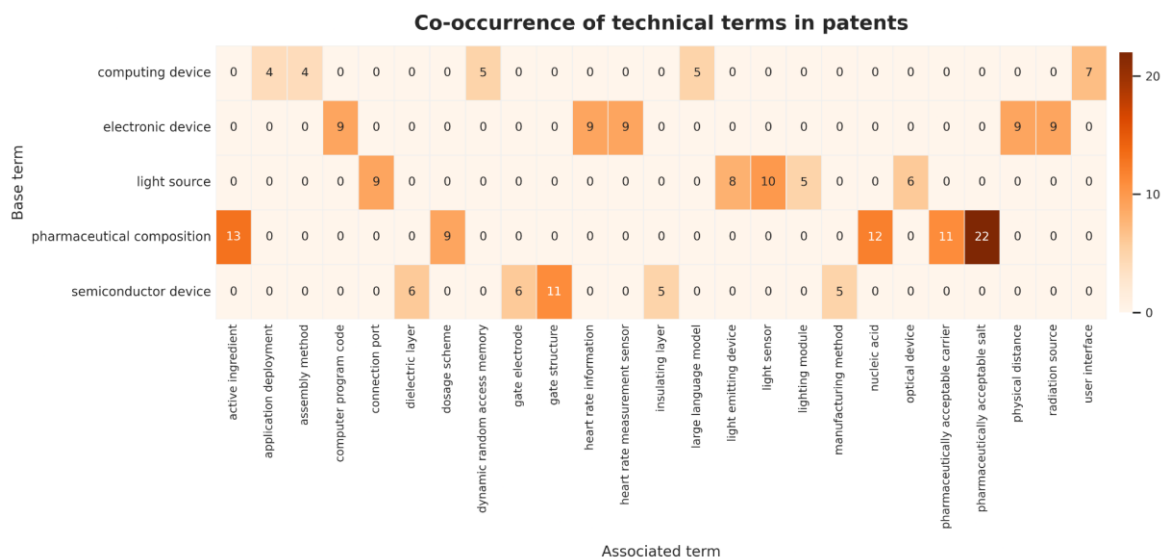


Figure 8 – Co-occurrence of technical terms in patents. Source: Prepared by the authors [2026].

In the semiconductor domain, the term semiconductor device shows the strongest association with gate structure (11), followed by dielectric layer and gate electrode, both with six occurrences. For light source, the most notable relationships are with light sensor (10), connection port (9), and light emitting device (8). The term electronic device is associated with terms such as computer program code, heart rate information, heart rate measurement sensor, physical distance, and radiation source, all with nine co-occurrences. Finally, computing device shows its strongest relationship with user interface (7), followed by dynamic random access memory and large language model, with five occurrences each.

These associations demonstrate how the multidimensional model expands the analytical possibilities of patent data. By integrating terms, documents, authors, countries, dates, and technological categories within the same structure, the Data Warehouse enables semantic relationships to be explored from different perspectives, combining temporal, thematic, and geographical dimensions. In this context, co-occurrence analysis complements thematic clusters by revealing specific relationships among terms that may be distributed across different groups. While K-Means organizes patent documents according to lexical similarity, the dimensional model makes it possible to examine these associations by period, category, country, or author, enriching the interpretation of the results. Since the values still correspond to raw counts, normalized measures such as Pointwise Mutual Information (PMI) may be incorporated in future analyses to identify more statistically representative relationships.

To address the limitations associated with isolated IPC letters, the technological categories were reconstructed at the IPC subclass level by concatenating the section, class symbol, and subclass fields available in the source data. This procedure produced 611 syntactically valid IPC subclasses and provided classification coverage for all 16,311 patents included in the dimensional database. Figure 9 presents the five most characteristic compound terms associated with each of the eight most frequent subclasses. Only compound terms occurring in at least three patents were considered. The terms were ranked using the thematic association score defined in Equation (1):

$$S_{c,t} = \left(\frac{100 \times n_{c,t}}{n_c} \right) \times \left(\frac{n_{c,t}}{n_t} \right) \times \log \left(\frac{n_t + 1}{n_t} \right) \quad (1)$$

where $S_{c,t}$ is the association score between subclass c and term t ; $n_{c,t}$ is the number of patents assigned to subclass c that contain term t ; n_c is the total number of patents assigned to subclass c ; n_t is the number of patents containing term t in the complete analytical sample; and N is the total number of patents in that sample ($N = 16,311$).

The first component measures the percentage coverage of the term within the subclass, the second measures its concentration in that subclass, and the logarithmic component corresponds to its inverse document frequency. The addition of one to the numerator and denominator of the logarithmic component prevents undefined values. Consequently, higher scores identify terms that are simultaneously recurrent within a subclass, concentrated in that technological category, and comparatively uncommon in the remaining patent collection.

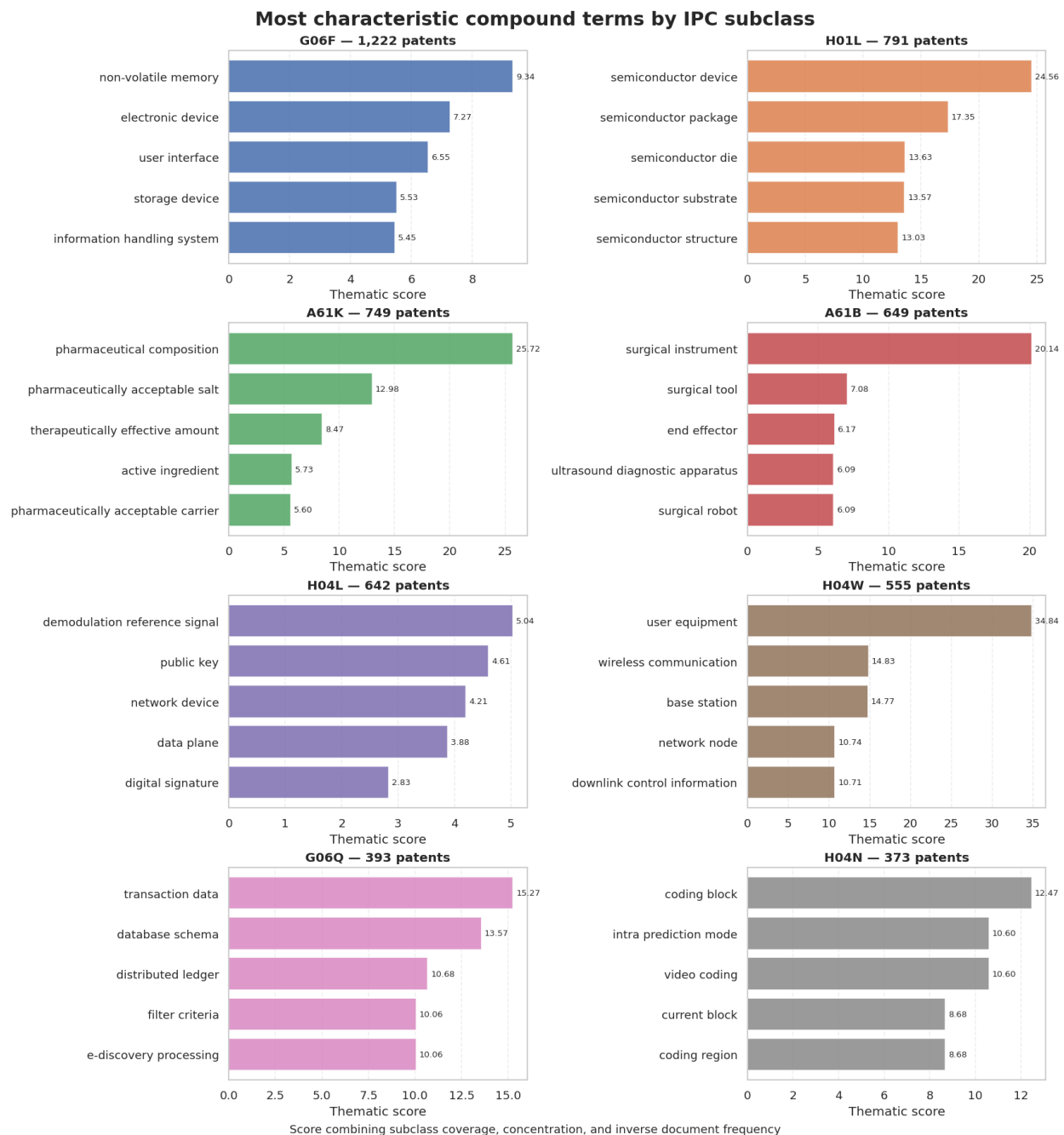


Figure 9 – Most characteristic compound terms associated with the eight most frequent IPC subclasses. Terms were ranked using a thematic association score combining their coverage within each subclass, their concentration in that subclass, and their inverse document frequency in the complete analytical sample. Only compound terms occurring in at least three patents were considered. Source: Prepared by the authors [2026].

The use of complete IPC subclass codes provides a more precise technological interpretation than the isolated alphabetic symbols adopted in the original analysis. G06F is the most frequent subclass, comprising 1,222 patents, and is characterized by terms related to digital data processing and storage, including non-volatile memory, electronic device, user interface, and storage device. H01L contains 791 patents and presents a consistent concentration of semiconductor-related expressions, such as semiconductor device, semiconductor package, semiconductor die, and semiconductor substrate. Similarly, A61K, with 749 patents, is distinguished by pharmaceutical composition, pharmaceutically acceptable salt, therapeutically

effective amount, and active ingredient, whereas A61B concentrates terms related to surgical instruments, diagnostic equipment, and surgical robotics.

The communication-related subclasses also present internally consistent vocabularies. H04L is characterized by expressions such as demodulation reference signal, public key, network device, data plane, and digital signature, while H04W contains terms related to wireless communication infrastructures, including user equipment, wireless communication, base station, network node, and downlink control information. G06Q is associated with transaction data, database schema, distributed ledger, and information-processing applications, whereas H04N concentrates expressions related to video coding and image communication. These results demonstrate that reconstructing the complete IPC subclasses eliminates the ambiguity produced by isolated letters and provides technological categories that are more consistent with the semantic content of the associated terms.

The multidimensional structure further allows these relationships to be analyzed according to periods, countries, authors, and other dimensions represented in the Data Warehouse. Therefore, the revised analysis not only improves the interpretation of the technological categories but also demonstrates how the dimensional model supports the exploration of relationships between structured patent classifications and technical terms extracted from patent documents.

4.5 Feasibility Demonstration of the MLP-Based Forecasting Module

Figure 10 illustrates the output of the MLP-based forecasting module for eight selected technical terms using normalized annual prevalence. For each term and year, the number of distinct patents containing the term was divided by the total number of documents available in that year and expressed per 1,000 documents. Each patent was counted only once for a given term, regardless of the number of times the term appeared in its textual content.

For visual clarity, the figure presents the most recent continuous observed window, covering 2022–2025. Solid lines represent the normalized rates observed during this period, whereas dashed lines represent the illustrative outputs generated by the MLP for the 2026–2028 horizon. The vertical dotted line separates the observed period from the illustrative projection period. The displayed values are intended solely to demonstrate the technical operation of the forecasting module and its integration with the dimensional database. They should not be interpreted as validated estimates of future technological activity.

The results demonstrate that the proposed architecture can retrieve term-level temporal data from the Data Warehouse, normalize them according to the annual number of available documents, organize the resulting rates as sequential inputs, and generate term-specific trajectories through the MLP component. Normalization reduces the influence of variations in annual corpus coverage and prevents repeated occurrences of the same term within a patent from disproportionately affecting the temporal series.

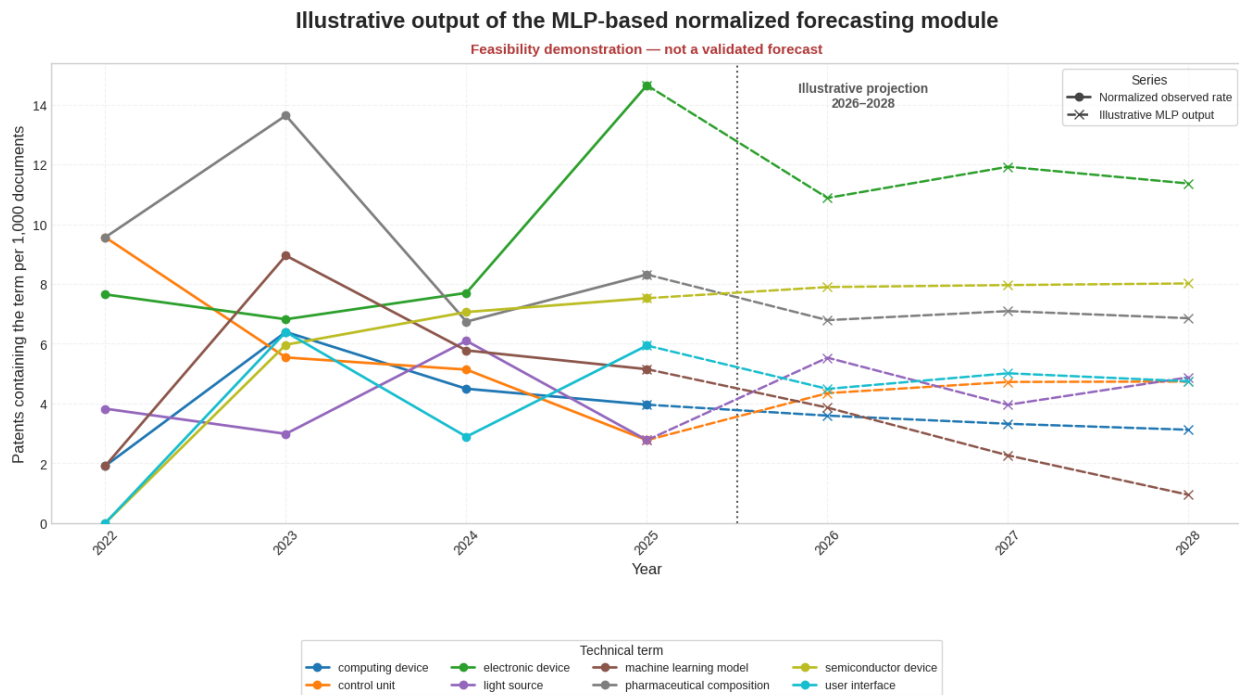


Figure 10 – Illustrative output of the MLP-based forecasting module using normalized annual prevalence for eight selected technical terms. The figure presents the observed period from 2022 to 2025 and the illustrative MLP outputs for 2026 to 2028. Values are expressed as the number of patents containing each term per 1,000 documents. Source: Prepared by the authors [2026].

Nevertheless, the limited length of the observed series and the incomplete monthly coverage of the source files restrict the predictive interpretation of the generated trajectories. The fluctuations displayed for some terms should therefore be understood as outputs of a feasibility demonstration rather than evidence of future technological growth, decline, or prevalence. A validated forecasting experiment will require a longer and continuous longitudinal dataset, a clearly defined temporal reference, chronological training, validation, and test subsets, evaluation through metrics such as MAE and RMSE, and comparison with naive and statistical baseline models.

5. Conclusion

This study presented an automated Data Warehouse architecture for collecting, transforming, organizing, enriching, and analyzing patent data. The implementation of the workflow, from XML file processing to dimensional model loading and data access through a RESTful API, confirmed the technical feasibility of integrating data engineering, text processing, multidimensional modeling, and artificial intelligence within a unified analytical environment.

The architecture transformed heterogeneous patent records into structured data suitable for analytical use. The pipeline incorporated cleaning, normalization, tokenization, stopword removal, inventor name standardization, temporal data processing, and AI-based semantic enrichment of patent titles and abstracts. Although the technical architecture was designed to support different AI providers, the semantic enrichment evaluated in this study was performed using the gpt-4o-mini model. This modular design facilitates the future substitution or comparison of language models without requiring changes to the remaining processing stages.

The use of a staging area, batch processing, and idempotency mechanisms enabled interrupted executions to be resumed without reprocessing previously enriched patents or duplicating records. The resulting dimensional model supported OLAP queries and analyses involving normalized temporal prevalence, emerging technical terms, term co-occurrence, IPC-subclass distribution, and thematic clustering. Temporal indicators were expressed in terms of patents containing a term per 1,000 documents, reducing the influence of differences in annual document coverage and repeated occurrences within the same patent. The MLP component was retained exclusively as a feasibility demonstration of the integration between the dimensional database and a forecasting module and should not be interpreted as a validated predictor of future technological activity.

The computational environment, database configuration, semantic-enrichment processing rate, token consumption, and execution times for representative analytical queries were documented to improve reproducibility. The results indicate that the proposed architecture can support the organization and retrieval of patent data and provide structured inputs for exploratory technological analyses. Nevertheless, the findings remain descriptive of the analyzed corpus and should not be generalized to worldwide patent activity.

The results must also be interpreted considering the limitations of the experiment. The number of available patent documents varied substantially across years, and some periods presented limited or incomplete coverage. Although the clustering configuration was reassessed using the elbow method, silhouette coefficients, and partition stability, the low silhouette values indicate limited separation among the resulting groups. The clusters should therefore be understood as predominant technological profiles rather than completely isolated categories. Similarly, the forecasting outputs were affected by the short and temporally irregular observed series and were not subjected to predictive validation.

Future work should expand the temporal and technological coverage of the database, extend semantic validation with independent domain experts, and complete the adjudication of divergent term classifications. Additional developments may include automatic classification by technological field, technical summarization of patent documents, lemmatization, named entity recognition, terminology normalization, and systematic comparison among language models. Forecasting experiments should employ continuous longitudinal data, chronological training, validation, and test subsets, error metrics such as MAE and RMSE, and comparisons with naive and statistical baseline models. Reproducibility may be further strengthened by maintaining a complete and versioned package containing the source code, database schema, container configuration, execution instructions, benchmark procedures, and a processed data sample.

References

- Almeida, E. C. (2011). Estudo de Viabilidade de uma Plataforma de Baixo Custo para Data Warehouse. arXiv preprint arXiv:1108.0729. Available at: <https://arxiv.org/abs/1108.0729>.
- Devlin, J. et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT 2019*, v. 1 (long and short papers), pp. 4171–4186.
- Golfarelli, M.; Rizzi, S. (2009). *Data Warehouse Design: Modern Principles and Methodologies*. McGraw-Hill.
- González, S. M.; Sakata, T. C.; Nogueira, R. R. (2020). NewsMiner: Enriched Multidimensional Corpus for Text-Based Applications. In: *Artificial Intelligence and Soft Computing – ICAISC 2020*. Springer, pp. 231–242.

- Inmon, W. H. (2005). *Building the Data Warehouse*. John Wiley & Sons.
- Joulin, A. et al. (2016). Bag of Tricks for Efficient Text Classification. arXiv preprint arXiv:1607.01759.
- Junior, V. F. et al. (2014). A Pesquisa Científica e Tecnológica. *Revista ESPACIOS*, v. 35, n. 09.
- Kimball, R.; Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. John Wiley & Sons.
- Kong, X. Y. (2015). *Arquitetura para Publicação de Dados Abertos em Sistemas de Governo: O Exemplo do SIASG*. Dissertação, Universidade Federal do Rio de Janeiro.
- Krestel, R. et al. (2021). A Survey on Deep Learning for Patent Analysis. *World Patent Information*, Elsevier, v. 65, p. 102035.
- Lin, T.-Y.; Chou, L.-C. A systematic review of artificial intelligence applications and methodological advances in patent analysis. *World Patent Information*, v. 82, 102383, 2025. DOI: 10.1016/j.wpi.2025.102383.
- Machado, G. V. et al. (2019). DOD-ETL: Distributed On-Demand ETL for Near Real-Time Business Intelligence. *Journal of Internet Services and Applications*, Springer, v. 10, n. 1, p. 21.
- Mikolov, T. et al. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
- Nogueira, R. R. (2020). *Business Intelligence na Prática: Modelagem Multidimensional e Data Warehouse*. 1. ed. UNIASSELVI, Indaial. ISBN 978-85-515-0454-3.
- Nogueira, R. R.; Gonçalves, A. L. (2025). Aspectos Técnicos e Tecnológicos Relacionados à Análise de Patentes: Uma Revisão Integrativa. In: *Anais do SIMPEX 2025 – Simpósio Internacional de Ensino, Pesquisa e Extensão do UniSENAI*.
- OECD. *OECD Patent Statistics Manual*. Paris: OECD Publishing, 2009.
- Pennington, J.; Socher, R.; Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In: *Proceedings of EMNLP 2014*, pp. 1532–1543.
- Radford, A. et al. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI Technical Report, San Francisco.
- Silva, J. V. F. S. (2023). *Comparação de Desempenho entre os Bancos de Dados PostgreSQL e Neo4j para Acesso a Dados Complexos*. Universidade Federal de Uberlândia.
- World Intellectual Property Organization — WIPO. *World Intellectual Property Indicators 2025*. Geneva: WIPO, 2025.
- Zhang, S. (2020). *Data Modeling and ETL with PostgreSQL*. Technical Report.