Abdelmalek Essaadi University

National School of Applied Sciences Al Hoceima

2nd Year Data Engineering

Project Report: Analysis of Patents in Virus Engineering

Prepared by:

Loubna boukayoua

Najma El boutaheri

Wanaim Essaadia

Elmzouri Fatima-zahra

Fatima Elzahrae El aissaouy

Framed by:

Professor Mrs. Annas Elhaddadi

College year: 2023/2024

Knowledge Thanks

We, the team comprising Loubna Boukayoua, Najma Elboutheri, Wanaim Essaadia, Fatima-zahra El mzouri, and Fatima Elzahrae El Aissaouy, would like to extend our heartfelt gratitude to Professor Mrs. Annas Elhaddadi and Professor Fadwa Bouhafer for their invaluable support and guidance throughout this project. Their expertise, encouragement, and dedication have been instrumental in the success of our work. We are truly thankful for the time and effort they invested in mentoring us, which has significantly contributed to our growth as aspiring engineers. This project has been a pivotal experience in our academic journey, equipping us with the skills and knowledge necessary to excel in our future careers.

Abstract

Virus engineering represents a cutting-edge field focused on the modification of viral genomes to enhance applications in gene therapy, oncolytic virotherapy, and vaccine development. This report delves into the patent landscape of virus engineering, leveraging big data techniques to extract valuable insights. The analysis identifies innovation trends, key players, technological advancements, and geographic distributions. By examining extensive patent datasets, the study aims to foster a deeper understanding of the field, aiding strategic decision-making and promoting future breakthroughs.

Table of Contents:

List of Figures:

Introduction

Virus engineering involves the innovative manipulation of viral genomes to develop new therapies and vaccines. By dissecting viral functions and optimizing delivery systems, researchers can create more effective treatments for various diseases. This field has garnered significant interest due to its potential in gene therapy, oncolytic virotherapy, and vaccine development.

Patents play a crucial role in this technological landscape, offering insights into innovation trends, key contributors, and technological advancements. Effective patent analysis not only helps in managing risks and making strategic decisions but also enhances a company's ability to innovate and compete.

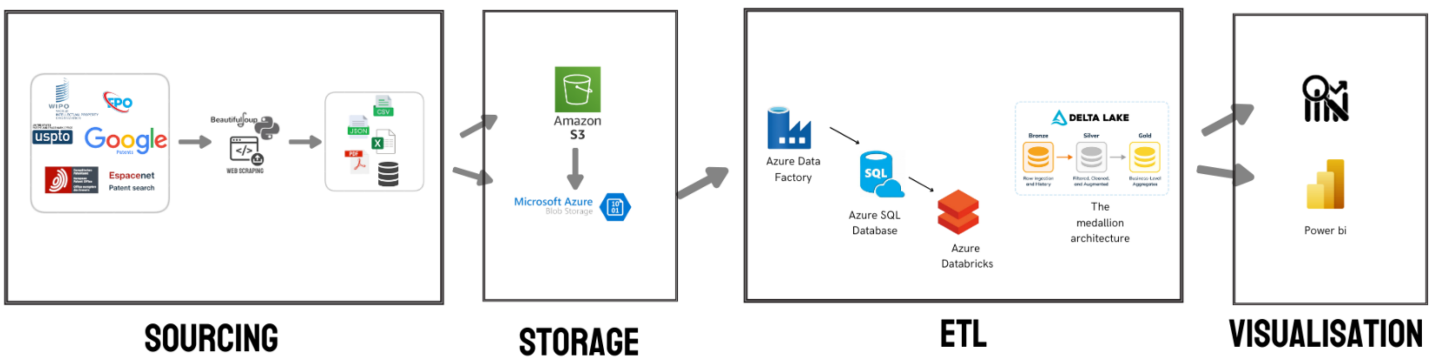
This report focuses on analyzing patents related to virus engineering using big data techniques. The comprehensive analysis encompasses data collection, cleaning, processing, integration (ETL), data warehousing, visualization, and in-depth analysis. By leveraging vast and complex datasets, this study aims to uncover trends, patterns, and insights within the field, ultimately fostering breakthroughs in gene therapy, oncolytic virotherapy, and vaccine development.

Objectives

The primary objective of this project is to conduct an in-depth analysis of patents in the field of virus engineering. By leveraging big data techniques, the study aims to:

- **Identify Key Trends:** Discover emerging trends and technologies in virus engineering from patent data.
- **Analyze Technological Advances:** Understand the innovations and breakthroughs documented in patents.
- **Identify Top Countries, Inventors, and Dates:** Highlight leading organizations, researchers, and inventors contributing to virus engineering.
- **Evaluate Geographic Distribution:** Analyze the geographic distribution of patents to identify regional strengths and collaborations.
- **Technology Landscape Analysis:** Provide a comprehensive overview of the technological landscape in virus engineering.
- **Competitive Intelligence:** Offer insights into the competitive dynamics within the field.
- **Innovation Trends:** Track the progression of technological innovations over time.
- **Prior Art Search:** Facilitate the identification of prior art to support future research and development.
- **Market Insights:** Provide market-related insights to guide strategic decision-making.
- **Technology Transfer and Licensing:** Explore opportunities for technology transfer and licensing to enhance commercial applications.
- Through this analysis, researchers and stakeholders can better navigate the evolving landscape of virus engineering, supporting the development of innovative solutions in gene therapy, oncolytic virotherapy, and vaccine development.

1. Architecture:



2. About Data and Soucring Process:

In this project, patent data was meticulously collected from various reputable sources using Python and the Beautiful Soup library. The data scraping process encompassed multiple websites known for their comprehensive patent databases, including Google Patents, Espacenet, Free Patents Online (FPO), the United States Patent and Trademark Office (USPTO), and the World Intellectual Property Organization (WIPO). A total of 35,500 patents were scraped, providing a robust dataset for thorough analysis.

- ### Google Patents:

Fields: The dataset from Google Patents includes detailed information such as ID, Title, Abstract, Description, Claims, Inventors, Current Assignee, Patent Office, Publication Date, and URL. This rich set of fields allows for an extensive examination of each patent's specifics, from the fundamental details to the broader context of the patent office and publication data.

- ### Espacenet:

Fields: Data from Espacenet covers Patent Number, Title, Inventor/Organization, Publication Date, Earliest Priority, Earliest Publication, and Description. Espacenet's focus on publication dates and earliest priority dates aids in understanding the timeline and priority of innovations within the field of virus engineering.

- ### Free Patents Online (FPO):

Fields: The fields scraped from FPO include Patent Title, Patent Number, and Abstract. While the dataset from FPO is more concise, it provides critical information for initial patent identification and categorization.

- ### United States Patent and Trademark Office (USPTO):

Fields: The USPTO data is the most detailed, with fields such as Document ID, Date Published, Description, CPCI (Cooperative Patent Classification), Inventor, Application Number, Application Name, Organization, Relevancy, IPC (International Patent Classification), Filing Date, and Family ID. This extensive dataset supports in-depth analysis, particularly regarding classification and relevancy.

- ### World Intellectual Property Organization (WIPO):

Fields: WIPO's dataset includes Patent Title, Patent Number, and Abstract. Like FPO, WIPO provides essential information for identifying and understanding the broader scope of each patent.

By utilizing these diverse sources and their respective data fields, the project ensures a comprehensive overview of the patent landscape in virus engineering. The integration of data from multiple websites allows for cross-referencing and validation, enhancing the reliability of the insights drawn from the analysis. The collected data is then processed, cleaned, and integrated, setting the stage for a detailed examination of innovation trends, key players, and technological advancements in virus engineering.

3. About Data: sourcing, preparation and understanding Workflow:

Our data storage and processing workflow employs Amazon S3 and Microsoft Azure Blob Storage for efficient data management. After sourcing data, we clean and preprocess it is using Apache Spark and Pandas, then transfer it to Azure for further tasks. Here's a detailed explanation of our storage and processing steps, along with data integration procedures.

a. Storage Workflow:

Initial Data Storage: **Amazon S3**

Data Sources: Data is sourced from various formats such as CSV, JSON, Excel, and PDF.

Storage in Amazon S3:

- **Scalability:** Ideal for storing large volumes of raw data.
- **Durability and Availability:** Ensures reliable and accessible storage.
- **Security**: Provides robust security features, including encryption and access control.

Example: Our Amazon S3 bucket contains diverse files like "espacenetpatents_data_v111.csv" and "fpoppatents1.csv."

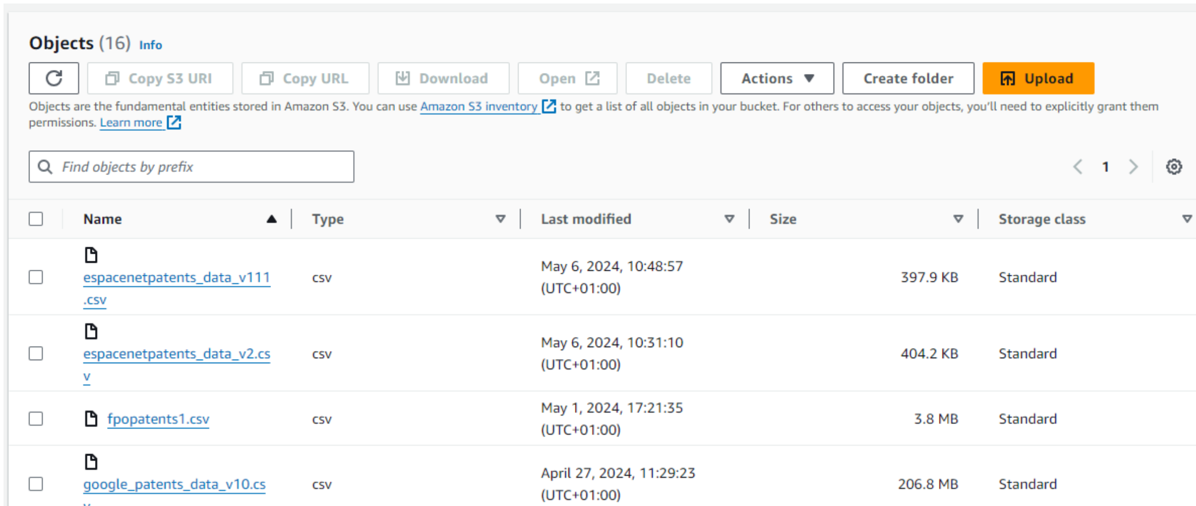


Figure 1: Files on S3 storage.

b. Data Preparation and Understanding:

The project leveraged a comprehensive approach to data preparation, sourcing patent information from multiple reputable databases. The data collected from these sources provided a robust foundation for analyzing the patent landscape in virus engineering. Here's an overview of the data sourced from each platform:

→Google Patents:

The dataset from Google Patents includes the following fields:

- **ID:** Unique identifier for each patent.
- **Title:** The official title of the patent.
- **Abstract:** A summary of the patent's content.
- **Description:** Detailed information about the patent's innovation.
- **Claims:** Specific legal claims made by the patent.
- **Inventors:** Names of the individuals who invented the patented technology.
- **Current Assignee:** The entity currently holding the rights to the patent.
- **Patent Office:** The office where the patent was filed.
- **Publication Date:** The date when the patent was published.

- **URL:** Link to the patent's webpage.

This rich set of fields allows for an extensive examination of each patent, offering insights into both the specific details of the inventions and the broader context of their publication and assignees.

→Espacenet:

Espacenet's data includes the following fields:

- **Patent Number:** Unique identifier assigned to each patent.
- **Title:** The official title of the patent.
- **Inventor/Organization:** Names of inventors or the organization holding the patent.
- **Publication Date:** The date when the patent was published.
- **Earliest Priority:** The earliest date of priority claimed by the patent.
- **Earliest Publication:** The earliest date the patent was made public.
- **Description:** Detailed information about the patented innovation.

Espacenet's focus on publication dates and earliest priority dates aids in understanding the timeline and priority of innovations within the field of virus engineering.

→Free Patents Online (FPO):

FPO provides a concise dataset with the following fields:

- **Patent Title:** The official title of the patent.
- **Patent Number:** Unique identifier assigned to each patent.
- **Abstract:** A summary of the patent's content.

Despite its conciseness, the FPO dataset provides critical information for initial patent identification and categorization.

→United States Patent and Trademark Office (USPTO):

He USPTO dataset is the most detailed, with the following fields:

- **Document ID:** Unique identifier for each document.
- **Date Published:** The date when the patent was published.
- **Description:** Detailed information about the patent's innovation.

- **CPCI (Cooperative Patent Classification):** Classification code used to categorize the patent.
- **Inventor:** Names of the individuals who invented the patented technology.
- **Application Number:** Unique number assigned to the patent application.
- **Application Name:** Name given to the patent application.
- **Organization:** Entity holding the patent.
- **Relevancy:** Measure of the patent's relevance.
- **IPC (International Patent Classification):** International code used to classify the patent.
- **Filing Date:** The date when the patent application was filed.
- **Family ID:** Identifier for the patent family, grouping related patents.

This extensive dataset supports in-depth analysis, particularly regarding classification and relevancy.

→World Intellectual Property Organization (WIPO):

WIPO's dataset includes the following fields:

- **Patent Title**: The official title of the patent.
- **Patent Number:** Unique identifier assigned to each patent.
- **Abstract:** A summary of the patent's content.

Like FPO, WIPO provides essential information for identifying and understanding the broader scope of each patent.

C. Data Storage Workflow: Azure Blob Storage

Initially, the raw data from these sources was stored in Amazon S3 (Simple Storage Service). Amazon S3 was chosen for its scalability, durability, and cost-effectiveness, making it an ideal solution for handling large datasets. After the initial storage, the data underwent cleaning and preprocessing to ensure quality and consistency. The cleaned and preprocessed data was then moved to Azure Blob Storage for further analysis and processing tasks. Azure's robust analytics and processing capabilities facilitated advanced data integration, analysis, and visualization.

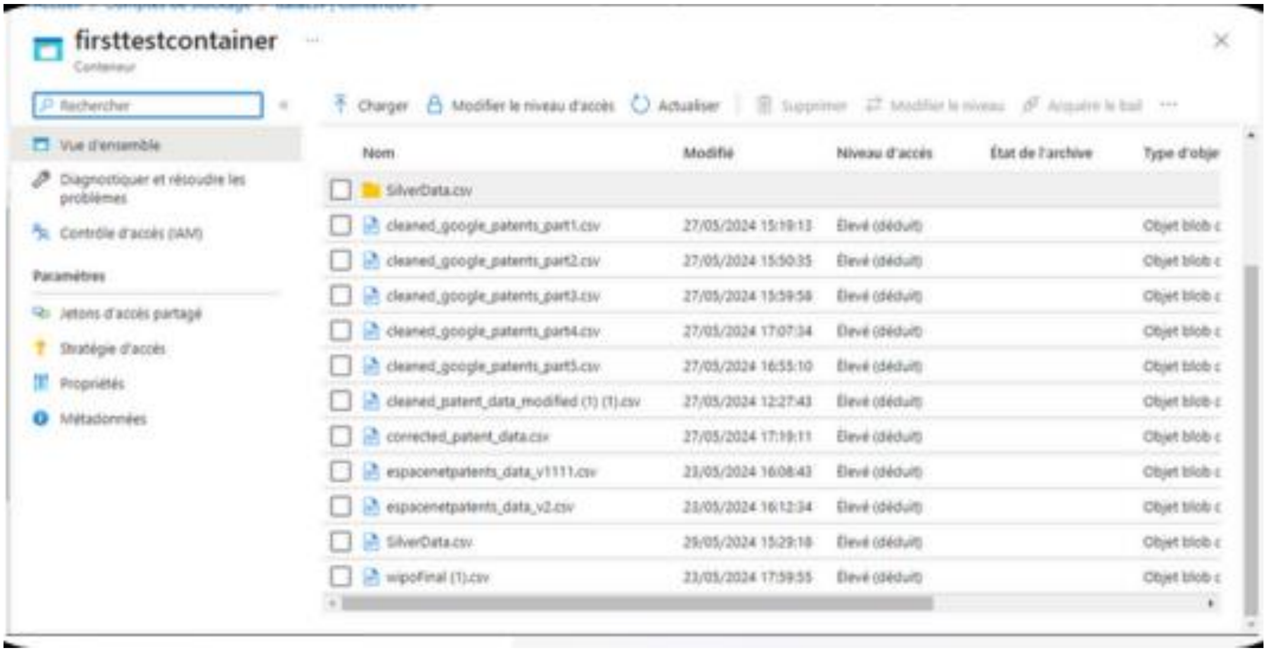


Figure 2: Files on Azure Blob Storage.

Summary:

By utilizing these diverse sources and their respective data fields, the project ensures a comprehensive overview of the patent landscape in virus engineering. The integration of data from multiple websites allows for cross-referencing and validation, enhancing the reliability of the insights drawn from the analysis. The collected data is then processed, cleaned, and integrated, setting the stage for a detailed examination of innovation trends, key players, and technological advancements in virus engineering.

4. ETL Process Using Azure Databricks:

The ETL process for patent data involves three main stages: extraction, transformation, and loading. Leveraging Azure Databricks, we implemented the Medallion architecture to manage the data flow efficiently.

a. Extraction

The extraction process begins with sourcing patent data from **AWS Storage.** We used **AWS** SDKs to connect and retrieve data, ensuring all relevant patent records were collected. This data was then transferred to Azure Storage using **Azure Data Factory**, providing a seamless transition between cloud environments. Then we used the same tool to ingest data into **Azure SQL Database**.

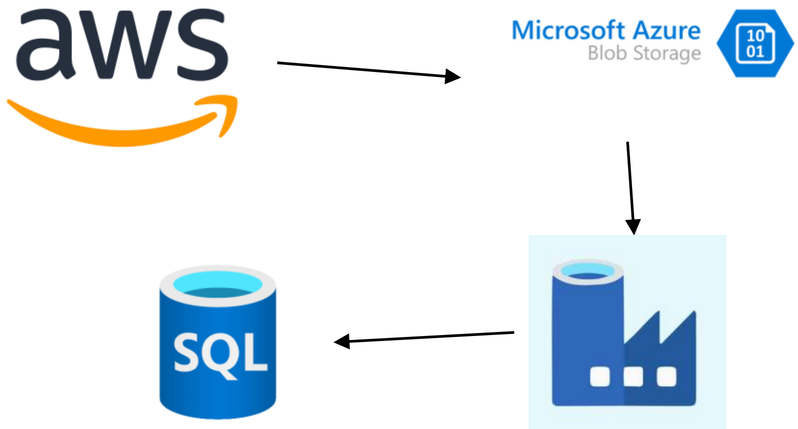


Figure 3: Data flow schema.

b. Transformation

In the transformation phase, we have followed the medallion architecture or the (**DELTA lake architecture)** to Improve data quality, we utilized **Azure Databricks** to process the data in three distinct stages:

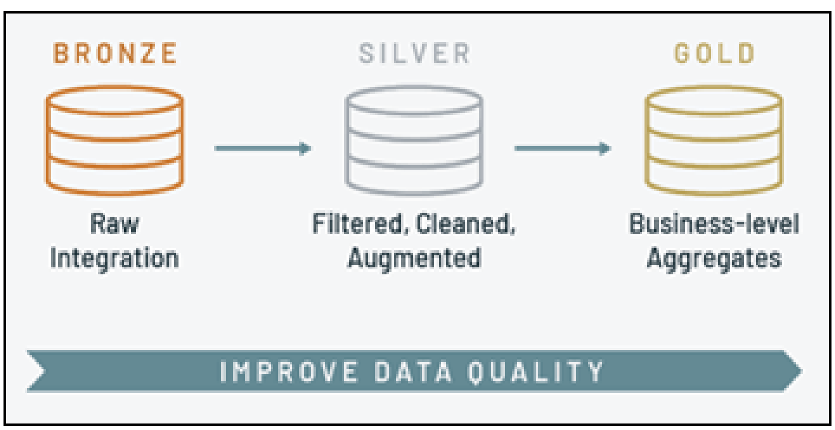


Figure 4: Delta lake architecture.

- **Bronze Stage**: Raw data was ingested into Databricks in its original format. This stage focused on reliable and reproducible data ingestion without any transformations, ensuring the raw data integrity.

- **Silver Stage**: Data cleaning and normalization were performed. This included filtering out irrelevant records, standardizing data formats, removing duplicates, and handling missing values. By the end of this stage, the data was structured and ready for more complex transformations.
- **Gold Stage**: The final stage involved aggregating and enriching the data. We applied business logic to summarize the data, joined it with external datasets for enrichment, and prepared it for analytical querying. This stage ensured the data was in its most refined form, ready for loading into the database

Figure 5: this image shows the bronze and gold table.

c. Loading

The transformed data was then loaded into the Azure SQL Database. We employed bulk loading techniques to ensure efficient data transfer and indexing strategies to enhance query performance. Post-loading, we performed data verification to ensure that the data's integrity and consistency were maintained throughout the process.

By following this structured ETL process, we ensured that the patent data was accurately extracted, comprehensively transformed, and efficiently loaded, ready for analysis and visualization in Power BI.

5. Power BI visualizations include:

→ Count of Patents by Year (Pie Chart):

This chart shows the distribution of patents across different years. Key observations:

- The year with the highest count of patents is 2017, accounting for 16.12% of the total.
- The next highest counts are in 2018 and 2016, with 15.24% and 15.81%, respectively.
- Other years have significantly lower counts, with some years like 2004 and 2003 having minimal representation.

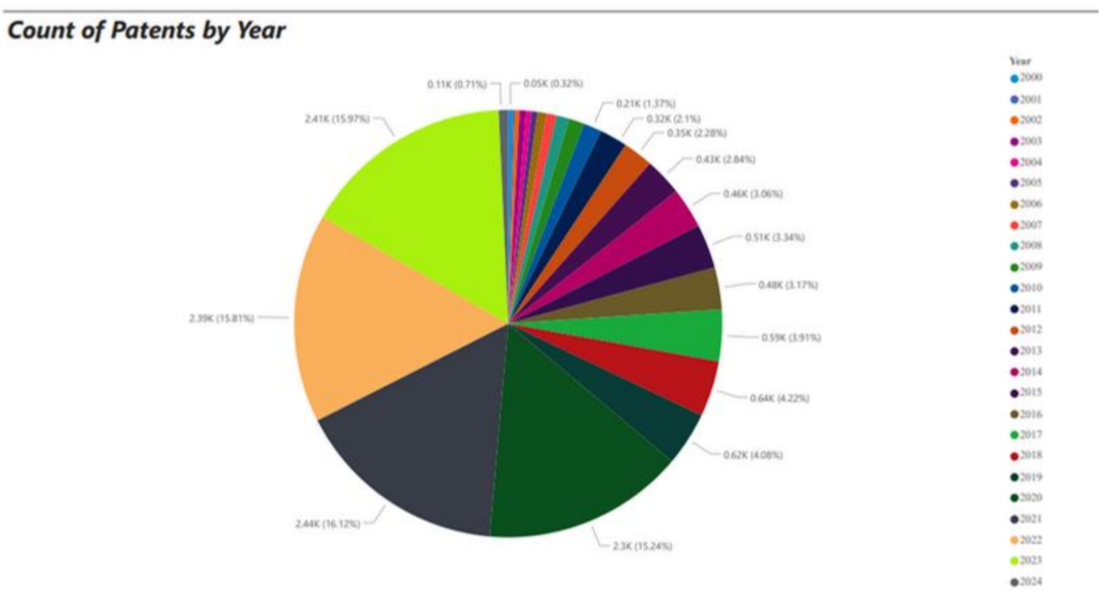


Figure 6: Pie chart shows count of Patents by Year.

→ Count of Used Languages (Bar Chart):

This chart displays the count of languages used in the data. Key observations:

- English (en) is overwhelmingly the most used language, far surpassing all others.
- The next most used languages, such as German (de), French (fr), and others, have significantly lower counts, barely visible on the chart.

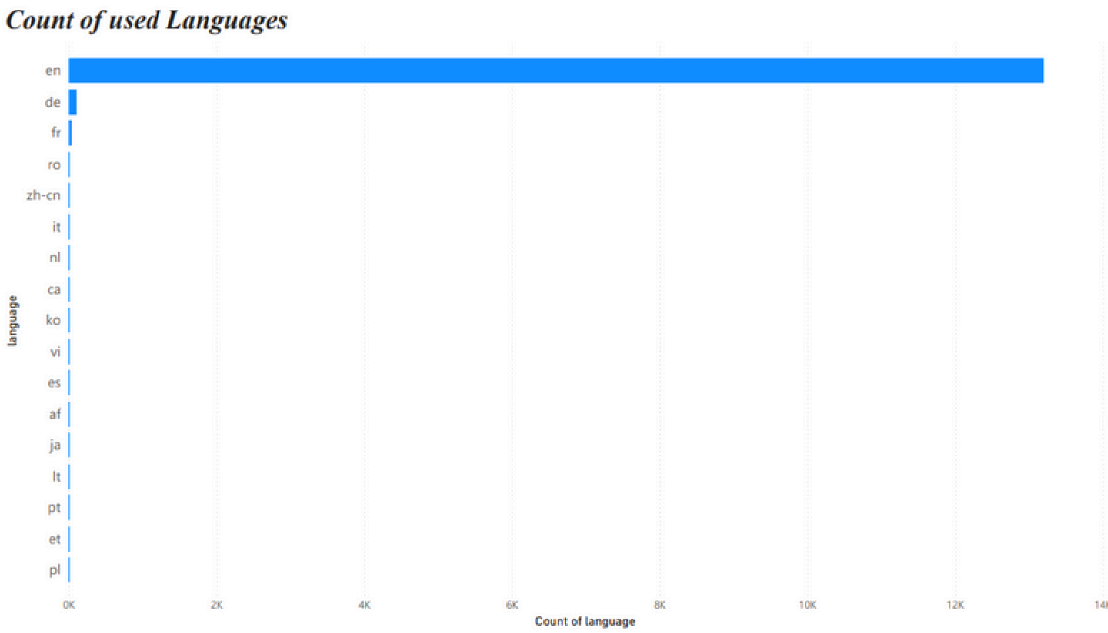


Figure 7: Bar Chart for Count of Used Languages

→ **Count of Inventors by Country (Bar Chart):** This chart shows the count of inventors by country.

→ **Key observations:**

China and the United States dominate the count, with China having the highest count followed closely by the United States.

Other countries like Australia, Russia, Canada, and Spain have much lower counts.

Many countries have minimal representation, indicating a highly concentrated distribution of inventors in a few key nations

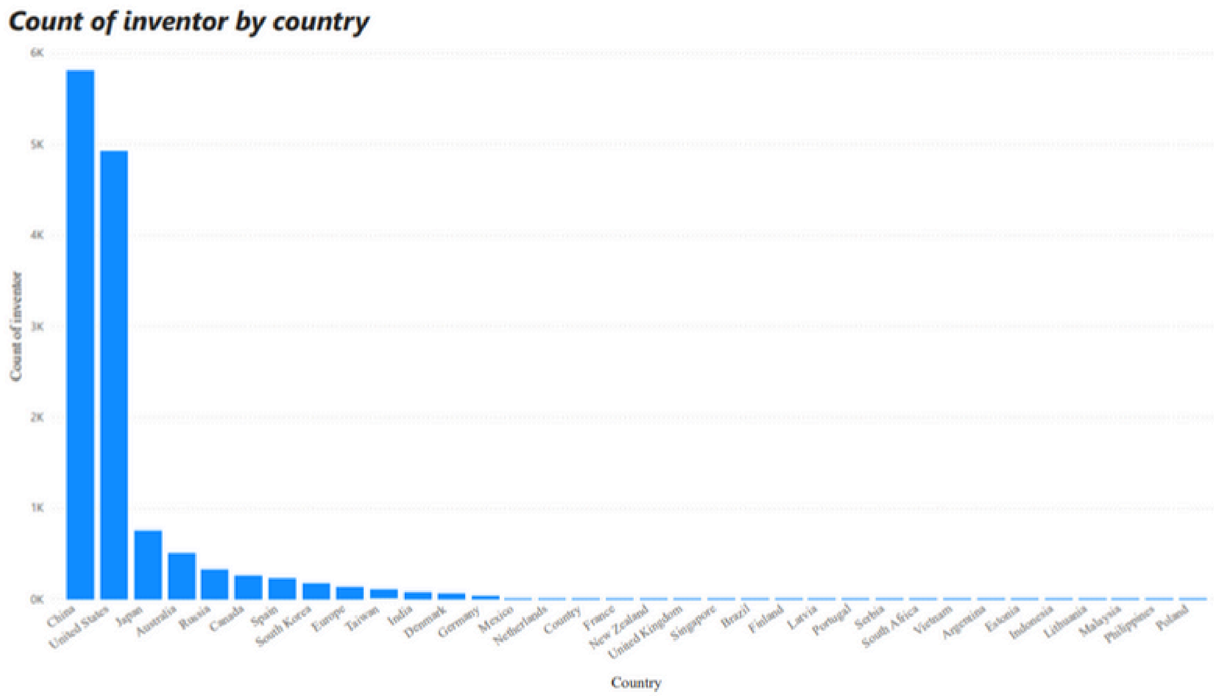


Figure 9: Chart of count of Inventors by Country

6. Web site:

A comprehensive overview of the website's functionality, design, and implementation. Here is a detailed outline of what is included in this section:

a. Introduction:

- **Purpose of the Website:** the primary purpose of the website is to allow users to search for patents related to virus engineering.
- **Target Audience:** the intended users are researchers, scientists, patent attorneys, and industry professionals.

b. Patent Search Functionality:

- **Functionality:** Users can select the source from which they want to retrieve patent data. The options include Espacenet, USPTO, WIPO, Google Patents, FPO, or All. The search bar allows users to refine their search based on their preference.

c. Search Results

- **Display Format:**
  - The search results are displayed in a tabular format with the following columns:
    - **Titre:** The title or brief description associated with each patent.
    - **Patent Number:** The unique number assigned to each patent.
    - **Inventeur/org:** The inventor or the organization that holds the patent.
    - **Date de publication:** The date when the patent was published or made public.
    - **Description:** A detailed description of the patented invention.
    - **Pays:** The country that granted the patent publication.
    - **Patent Language:** The language in which the patent is published.

- **Example Table:**

Figure 10: The search results are displayed in a tabular format.

Patent results

ID	Date	ABSTRACT	Inventor
C07K 14/00	06.10.2022	The present invention relates to a virus vaccine based on virus surface engineering, wherein the virus vaccine provides increased immunity. A linker peptide according to one aspect of the present invention has the characteristic of being attachable to a virus, and thus can improve the immunogenicity of a vaccine by being used as a linker capable of effectively binding, to the virus surface, an immune-enhancing substance that activates the immune system. When the linker peptide is combined with virus surface engineering technology, the immune-enhancing substance can be attached to the virus surface, and thus the linker peptide may be effectively used as a vaccine platform providing increased immunity.	SHIN, Hyun-J
C07K 19/00	13.05.2015	The objective of the invention is to provide a duck tembusu virus genetic engineering subunit vaccine. According to the invention, biotechnology is employed for antigenic epitope analysis and splicing of a duck tembusu virus E protein so as to obtain a novel duck tembusu virus fusion protein DE; the obtained fusion protein is used as an antigen for preparation of the duck tembusu virus genetic engineering subunit vaccine; the amino acid sequence of the coded protein of the novel duck tembusu virus fusion protein DE is SEQ ID No.1; one of the nucleotide sequences of the coded protein of the novel duck tembusu virus fusion protein DE is SEQ ID No: 2; a	HOU ZHUME

- **Technologies Used**
- **Front-end**

- **Technologies and Frameworks:**
  - **HTML:** For structuring the web pages.
  - **CSS:** For styling the web pages and making them visually appealing.
  - **JavaScript:** For adding interactivity and dynamic elements to the web pages.

- **Back-end**

- **Technologies:**
  - **Node.js:** Used for building the server-side logic and handling requests from the front-end.
- **Data Storage:**
  - **CSV Files:** Patent data is stored in CSV files, which are read and processed by the back end to provide search results.

d. Example Workflow of the Website

- **User Registration and Login:**
  - **Sign-up Page:** New users can create an account by providing their email, creating a password, and filling in additional details as required.
  - **Login Page:** Registered users can log in using their email and password.

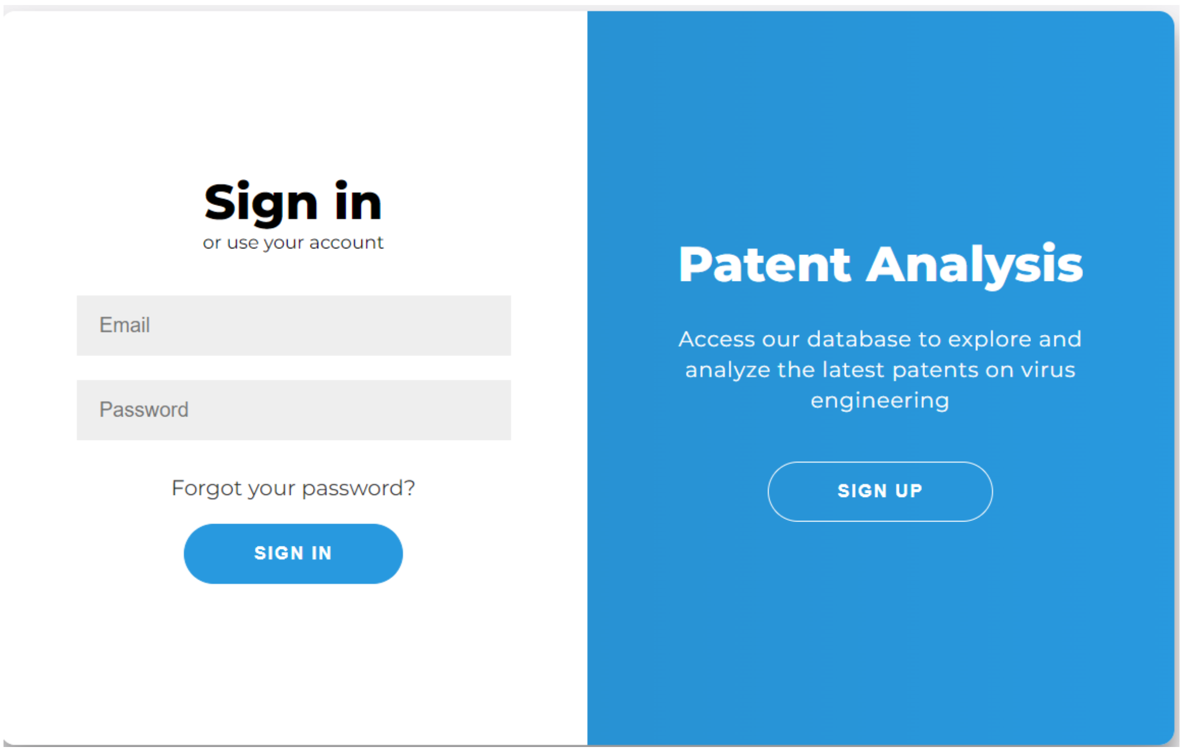


Figure 11: User Registration and Login

- **Performing a Search:**

➢ **Select Source:** The user selects "All" from the dropdown menu to search across all databases (Espacenet, USPTO, WIPO, Google Patents, FPO).

➢ **Submit Search:** The user clicks the search button, and the website processes the request.

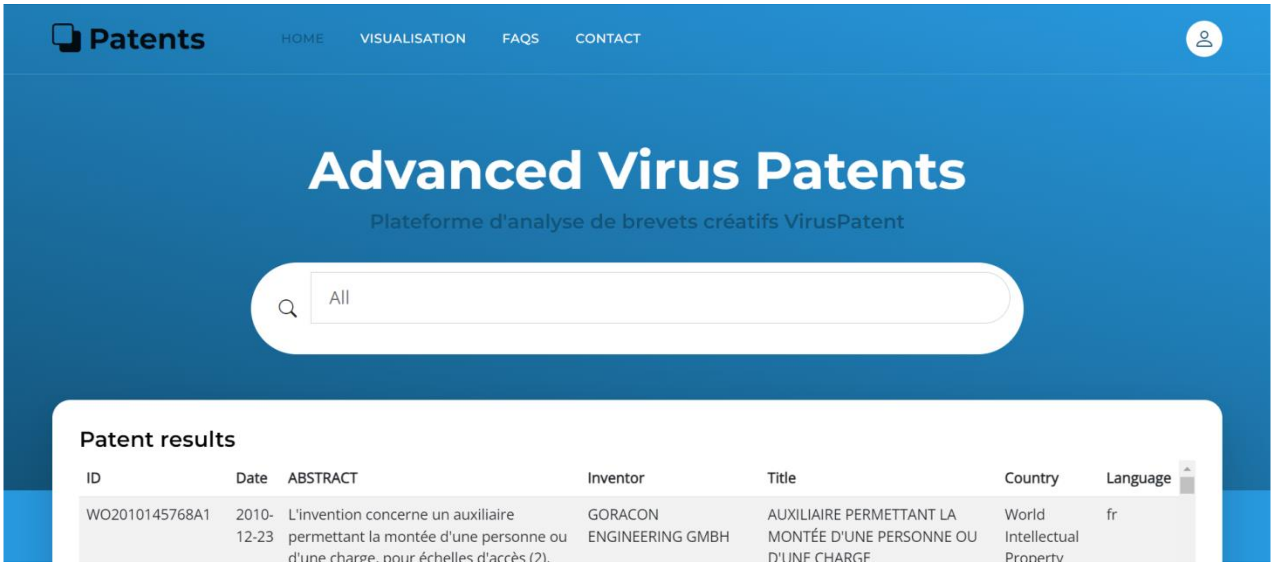


Figure 12: User Registration and Login

- **Displaying Results:**
  - **Fetch Data:** The back_end fetches relevant patents from the selected sources.
  - **Parse Data:** The data from CSV files is parsed and organized into a structured format.
  - **Display Table:** The front-end displays the search results in a table with the specified columns.
- **Data Visualization:**
- **Purpose:** To provide users with visual insights into the patent data, making it easier to identify trends, patterns, and key information.
- Types of Visualizations: Bar Charts and Pie Charts.

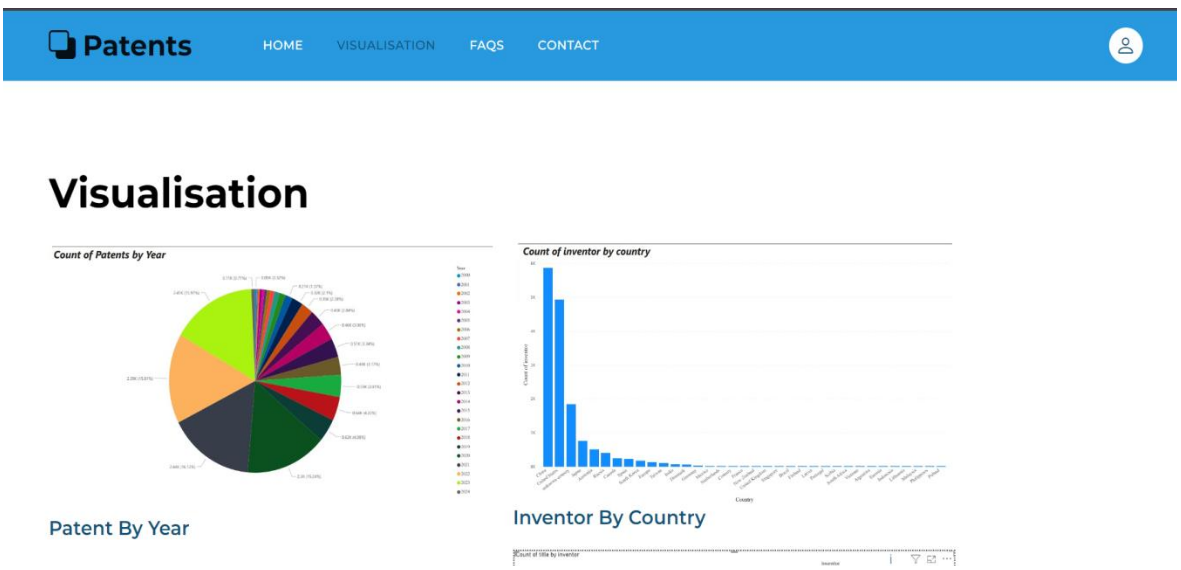


Figure 13: Data Visualization

- **Frequently Asked Questions (FAQ) Section**

To provide users with quick answers to common questions about patent website, ensuring a betteruser experience and reducing the need for direct support.

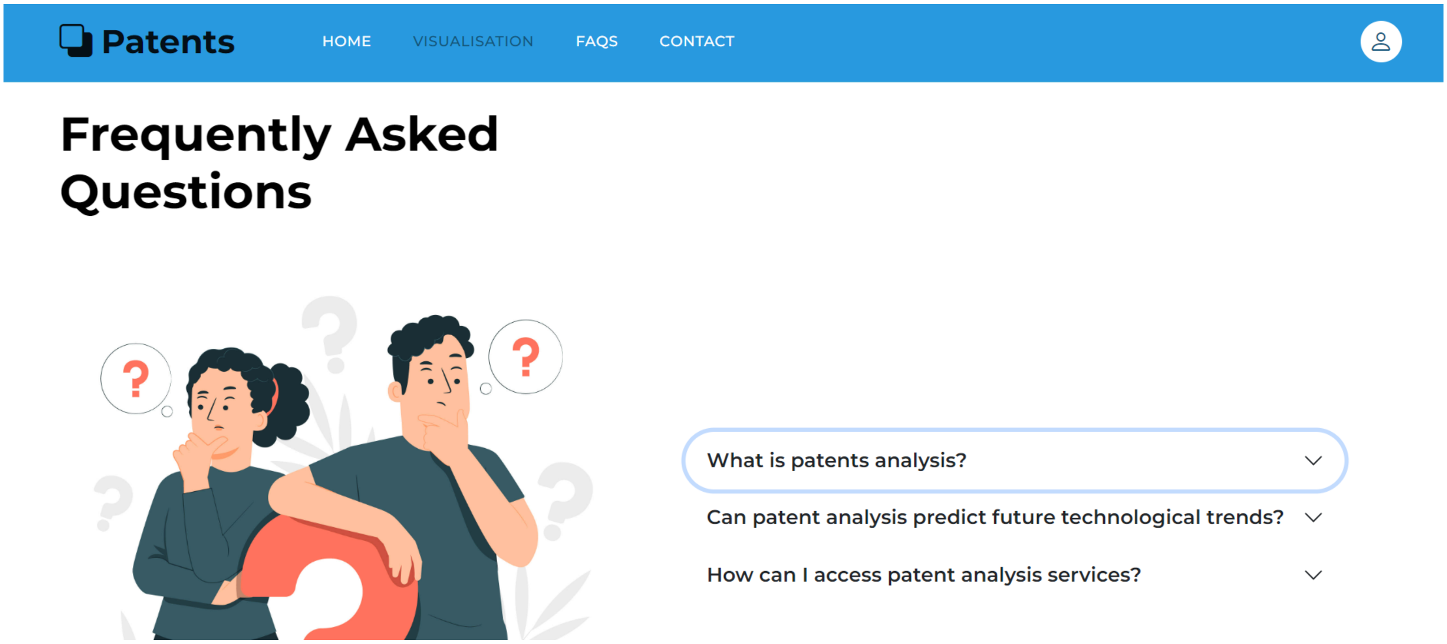


Figure 14: Asked Questions

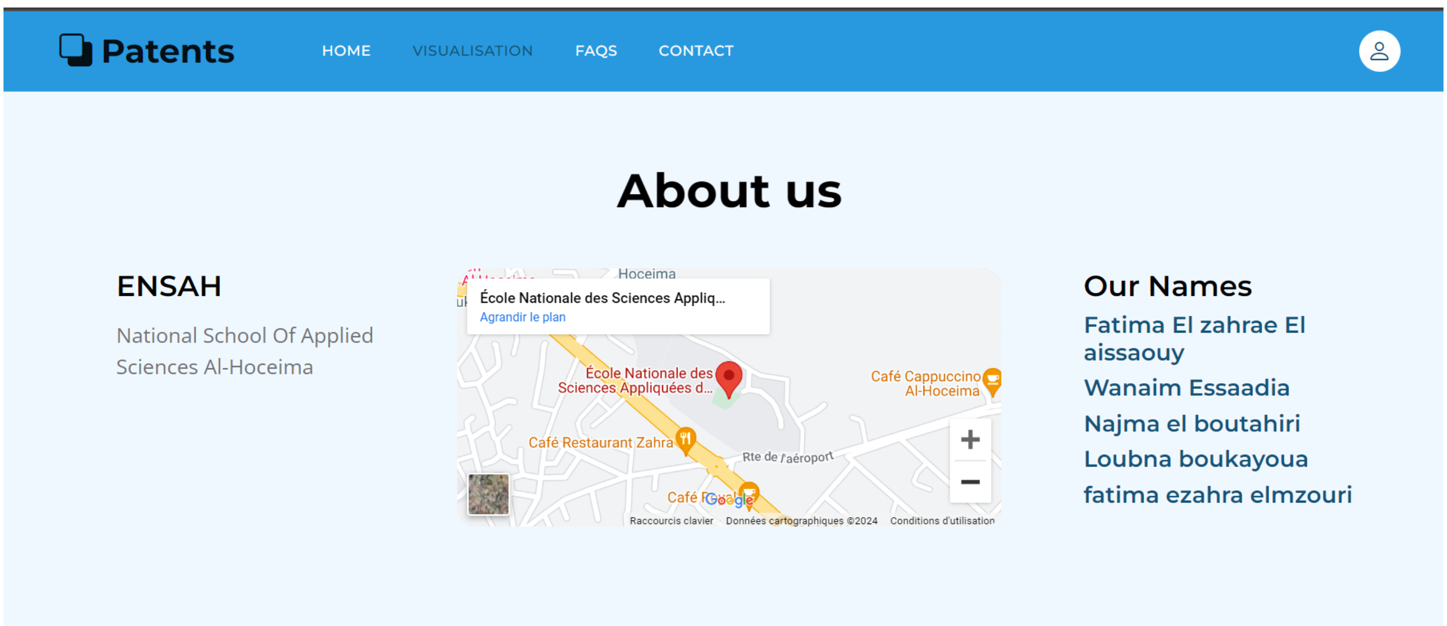


Figure 15: About us

Conclusion

The "Analysis of Patents in Virus Engineering" project has been a comprehensive and enlightening endeavor. From the initial stages of data sourcing, through storage and data warehousing, to the final visualization and analysis, this project has provided us with invaluable hands-on experience in the field of data engineering and analytics. Throughout this project, we developed critical skills in data management, big data technologies, and analytical techniques, which are essential in the rapidly evolving field of virus engineering. This project not only deepened our understanding of technological advancements and innovation trends within the domain but also prepared us for future professional challenges by enhancing our technical expertise and problem-solving abilities

Summary

This report focused on analyzing patents related to virus engineering using big data techniques. The analysis aimed to uncover trends, patterns, and insights within the field, ultimately fostering breakthroughs in gene therapy, oncolytic virotherapy, and vaccine development. The study involved comprehensive data collection, cleaning, processing, integration, data warehousing, visualization, and in-depth analysis.

✓ Key Insights and Results:

- Identification of Key Trends: The study discovered emerging trends and technologies in virus engineering from patent data, highlighting the increasing focus on gene editing techniques and novel delivery systems.
- Technological Advances: Innovations and breakthroughs documented in patents were analyzed, revealing significant advancements in oncolytic viruses and vector optimization for gene therapy.
- Top Contributors: The analysis highlighted leading organizations, researchers, and inventors contributing to virus engineering, with notable contributions from institutions in the United States, China, and Europe.
- Geographic Distribution: The geographic distribution of patents indicated regional strengths and collaborations, with a high concentration of patents in North America and Asia.
- Technology Landscape Analysis: A comprehensive overview of the technological landscape was provided, showing a dynamic and rapidly evolving field with significant potential for future developments.
- Competitive Intelligence: Insights into the competitive dynamics within the field were offered, identifying key players and their strategic positions.
- Innovation Trends: The progression of technological innovations over time was tracked, demonstrating a continuous increase in patent filings and technological advancements.
- Market Insights: Market-related insights were provided to guide strategic decision-making, emphasizing the growing commercial potential of virus engineering technologies.
- Technology Transfer and Licensing: Opportunities for technology transfer and licensing were explored, highlighting pathways to enhance commercial applications and industry collaborations.

✓ Future Work:

Building upon the foundation laid by this project, several avenues for future work can be pursued to further advance the field of virus engineering:

- Expanded Patent Dataset Analysis: Increase the scope of the patent dataset to include more comprehensive global data and cover a broader range of subfields within virus engineering.
- Machine Learning Integration: Implement advanced machine learning techniques to enhance the accuracy and depth of patent analysis, enabling predictive insights and trend forecasting.
- Collaborative Network Analysis: Conduct an in-depth analysis of collaborative networks among researchers and institutions to identify key partnerships and their impact on innovation.
- Real-time Data Integration: Develop systems for real-time integration and analysis of new patent filings to provide up-to-date insights and maintain a competitive edge in the field.
- Commercialization Pathways: Explore detailed pathways for commercialization of key technologies identified in the patents, including potential market entry strategies and business models.
- Policy and Regulatory Impact: Analyze the impact of policy and regulatory changes on the patent landscape to understand their influence on innovation and technological development.
- Cross-disciplinary Innovations: Investigate cross-disciplinary innovations by integrating insights from related fields such as biotechnology, nanotechnology, and computational biology to drive new breakthroughs.
- By pursuing these future directions, researchers and stakeholders can continue to navigate and shape the evolving landscape of virus engineering, supporting the development of innovative solutions in gene therapy, oncolytic virotherapy, and vaccine development.

Bibliography and Webography:

» https://patents.google.com/?q=(virus+enginner)&oq=virus+enginner
» https://worldwide.espacenet.com/patent/search/family/036617365/publication/JP2019146560A?q=virus%20enginner
» https://www.wipo.int/tools/en/gsearch.html?cx=01645853759490540650 6%3Ahmturfwvzzq&language=en&allinurl=wipo.int%2Fwipolex&hl=en&lr=&cof=FORID%3A11&tab=1&q=virus+enginner+#gsc.tab=0&gsc.q=virus%20enginner%20&gsc.page=1
» https://www.uspto.gov/patents/search
» https://www.freepatentsonline.com/#google_vignette