Final Report
Computational Spatial Humanities

Industry Change Represented By Large Corporate Headquarters

Jiacheng Lang, Jonas Greim
Lecturer: Dr. Thomas Efer

October 15, 2024

Contents

1 Introduction

Over the past several decades, research on the location of company headquarters has been an ongoing area of study. The studies [Bur77; SH81; HW91; KT02], researched on tracking patterns of centralization - decentralization of headquarters. From the 1950s to the 1970s, in aspects of centralization, researchers found that large cities consistently had a more substantial capacity to accommodate corporate headquarters than smaller cities due to their ability to provide sufficient professional infrastructure, public services and resources. In aspects of decentralization, there was also a trend of headquarters dispersing from traditional industrial centers like New York [Bur77] and shifting from city centers of the metropolitan areas to suburban regions of the areas [SH81].

Since 1967, agglomeration theory has been continuously developed and applied as a significant theory in the study of spatial patterns of headquarters. Mills emphasized [Mil67] the heterogeneity of land and the impact of transportation costs on urban structure, how transportation costs and land structure impact urban organization. Increasing returns to scale was an essential factor in urban agglomeration, with spatial proximity facilitating the spillover of technology and information. This effect applied not only to specific activities but also to vertically related activities. Porter developed Mills' agglomeration theory. Porter argued [Por98] that location influences competitive advantage by rising productivity and productivity growth, not only by factors of production (e.g., land, transportation, natural resources, capital, labor). The importance of industrial cluster, a critical mass of companies in a particular sector in a particular location, in boosting productivity (value created per day of work or unit of capital) and fostering innovation is highlighted. Porter further developed the idea that location-specific advantages remain crucial in a globalized context, e.g., specialized skills, knowledge, and institutions. Emerging high-tech firms benefit from agglomeration and functional integration, while firms that rely on rapid product differentiation to meet consumer demands gain from proximity to competitors. In the new competitive landscape, firms succeed not by improving production efficiency alone but by launching new products and responding to market demands more quickly than their competitors.

Other studies have applied agglomeration theory to their analyses. The study of Shilton [SS99] supported agglomeration theory, highlighting a shift in focus from production efficiency to productivity improvements and summarizing different agglomeration characteristics by different industries. Another study by Davis [DH08] examined the agglomeration of intermediary service suppliers, finding that the diversity of local service providers and the presence of other nearby headquarters strongly impacted headquarters agglomeration. Other studies include the study [ARY91] on the impact of corporate headquarters relocation on stock prices, and the study [Gar+02] worked on inter-city competition on tax policies to company headquarters relocation.

Our study utilized the Scrapy web crawling framework to crawl the company data. We used the company name in the ranking list through Wikidata API to get the corresponding location data. We manually corrected the data to create two datasets to study the change in large corporation headquarters locations in the U.S. These datasets cover the companies of the Fortune 500 from 1958 to 2005 and those of the S&P 500 from 1990 to 2024. The metadata in these datasets includes company names, corporate Wikidata IDs, time, location coordinates, and industry categories. Our dataset has a larger time scale than the previous research, allowing us to track recent migration trends further. On our website, we provide time series data for the company headquarters of large companies, including each company's industry category, ranking, revenue, and profit in the Fortune 500 dataset and industry category, ranking, and market cap in the S&P 500 dataset. Additionally, we offer the distribution of headquarters by industry for each year. Through the spatial visualization of the time, location, and industries of large company headquarters, we aim to address the following questions:

1. Is the trend of company headquarters moving out of the Northeastern U.S. still ongoing?
2. In the digital and globalization era:
   a) has technology and globalization led to decentralization of headquarters location, or companies has continued to cluster their headquarters to leverage the synergies of agglomeration in rising competitiveness and addressing the growing complexity of competition?
   b) Does the headquarters location reflect a shift in company business strategy from cost minimization to productivity enhancement?
3. How has the industry changed? How has the ability of different regions to attract and accommodate headquarters from different industries changed?

2 Literature Research

This section outlines the literature research of the study, covering the definition of headquarters, the significance of headquarters, factors influencing headquarters location, history of headquarters location, and the reflection of headquarters location on economy and industry change.

2.1 Definition Headquarters

From the outside of the company, the company faces external competition and obtains information from the outside. From an internal point of view, enterprises need to deal with production, sales, management, decision-making, and other affairs. Companies position their headquarters as command and control centers to respond to this competition [GS92]. To face the competition, management varies the range and concentration of its products and services to improve the firm's development, profit, and pricing [Por98].

2.2 Significance of Headquarters

A company's headquarters is essential as it is the central hub for operations and strategic decision-making, significantly influencing its identity and performance. The headquarters functions as a command center where critical business strategies are developed and executed, housing top executives who shape the company's vision and culture.

Moreover, the headquarters symbolizes the brand and reputation of the firm, affecting its prestige and market position, which in turn influences stakeholders and customers. A well-placed headquarters enhances visibility and attractiveness to potential clients and talent, strengthening the company's competitive edge.

Additionally, it is a focal point for innovation and collaboration, facilitating communication and coordination among various departments. This centralization streamlines operations and enhances the ability to respond swiftly to market changes. A headquarters's strategic location and functionality are vital for a company's long-term growth and sustainability in a competitive landscape.

2.3 Factors Influencing Headquarters Location

The definition of the headquarters location means the cities of urban centres where the offices are found [Ahn84]. The location of corporate headquarters is a complex and critical decision-making process influenced by a wide range of factors. The following are some key factors affecting the location of company headquarters, covering the establishment of companies, human resources, government policy, and population.

2.3.1 The Establishment of Companies

The first thing to clarify is that the location of the headquarters of many businesses is determined by where the business was located when it was established. Shifts in headquarters dominance by the city are related less to the headquarters relocation than to the growth of local companies. When the local companies grow large enough to be included in the Fortune 500 list, the large corporation headquarters dominance of the city is also enhanced [HW91].

Three relationships exist between the company's establishment and headquarters location [Ahn84]. The first type is that the founders started the business there, and no specific location decision exists. The second type is the location decision without concerning the headquarters function. These locations are dedicated to manufacturing activities that are beneficial for transporting and acquiring raw materials, e.g., Volkswagen's headquarters location $Wolfsburg$. The third type is location decision concerning the headquarters function, the purpose of which is centralizing management, coordinating "multi-locational" operations, and maintaining essential face-to-face connections. Essentially, it is about considering how the headquarters contributes to the overall functioning and success of the company and having a bird-eye view and strategic thinking on the whole business.

2.3.2 Human Resource and Information Acquiring

The labor market plays a crucial role in the location choice of enterprises. In the post-industrial era, whether a region can provide enough well-educated labor, mainly technical and managerial talents, directly affects the location decisions of enterprises. Second, differences in wage levels in different regions also affect employees' operating costs and quality of life, affecting a firm's ability to attract and retain talent. Firms that locate their headquarters near universities and research institutions demonstrate a reliance on technical and specialized talent, which contributes to the growth of high-skilled jobs in the local economy. At the same time, headquarters tend to be concentrated in high-income areas, reflecting these areas' high standard of living and purchasing power, which can attract and retain a well-qualified labor force. The technological ecosystem created by clustering high-tech firms in certain areas can attract startups or technology companies that choose to locate there. This clustering effect provides enterprises with rich opportunities for collaboration and a resource-sharing platform, promoting innovation and development. In addition, firms sometimes choose to locate close to competitors in order to obtain market information and technological inspiration to enhance their competitiveness.

2.3.3 The Policy of Government

The economic environment is one of the fundamental factors affecting the company headquarters location. The Government must strive to create an environment that supports rising productivity [Por98]. The appropriate government policy in fields, e.g., antitrust, intellectual property protection, and tax breaks, can encourage the development of local industries. Not all companies can provide positive externalities to a specific location regarding tax breaks. Governments are more willing to offer tax breaks when a company brings potential benefits to the local economy [Gar+02]. Thus, a government's tax policy towards companies is tailored according to the local economic structure, with different policies for different companies. Even the same company may receive different government policies in different locations. The Government policy also can remove obstacles and constraints. The Government cannot only set supply-side policies but also demand-side policies to use the local market to help grow new products and encourage innovation.

2.3.4 The Change of Population

The population has increased in the South and West of U.S. after the end of the second world war. Population growth in the South and West means that consumer markets in these regions are also expanding rapidly. Relocating corporate headquarters to these regions allows for better proximity to customers and markets, increasing sales and business opportunities. For historical reasons, the traditional labour unions in the southern and western states were also weak. A large number of new population provided a relatively cheap and high-quality labour force, forming the centre of the energy industry in Texas in the southern part of the U.S. and the centre of the technology industry in Silicon Valley in the western part of the country.

2.4 History of Headquarters Location

A study targeting the period from 1955 to 1975 showed [SH81], the headquarters of U.S. industrial firms experienced a degree of decentralization and formed a more geographically balanced distribution. More headquarters located in the emerging South and West of U.S. The headquarters relocation was obviously relative to the industry change and the transformation from an industrial society to a post-industrial one.

The study [HW91] emphasized that the deconcentration of the headquarters location continued in the 1980s. Four of the five most significant metropolitan areas in the U.S. suffered a noticeable decline in the number of headquarters of Fortune 500 corporations and assets control. The mainstream trend of the 1970s is the decline of the frost-belt and the growth of the sun-belt. This trend could have been more evident in the 1970s compared to the 1980s.

In the 1990s, metropolitan areas were still the most attractive headquarters locations [KT02]. The study mainly studied the situation of headquarters locations in cities of the same level and found that headquarters were shifting from first-tier to second-tier cities. Large cities in the South were the biggest winners, while cities on the West Coast stayed relatively behind in the 1990s.

2.5 How Headquarters Location reflects Economy and Industry Change

This section outlines the reflection of the headquarters location on two aspects: the U.S. economy and industry change.

2.5.1 Headquarters Location reflects Economy

Most headquarters may be there because the city was historically a firm breeding ground [Ahn84]. The presence of numerous headquarters in a particular town should not be viewed simply as a result of location choices that leverage its qualities as a headquarters. Due to the city's historical reputation as a fertile business ground, many of these offices are likely situated there. The town provided service infrastructure, ample financial and market opportunities, allowing many companies to flourish and eventually acquire others established in different locations. As these businesses expanded into multiple locations, they recognized the need to centralize their administrative and control functions at headquarters. These headquarters were established in the original city because that is where the parent companies of the growing corporations were based. The founders in these companies were often reluctant to relocate, resulting in a network of essential face-to-face connections. The concentration of headquarters is mainly due to the city's capacity to nurture new businesses. This developmental potential created appealing that motivated decision-makers to choose the city as their headquarters location. Thinking about the large corporations' headquarters locations from this perspective, we can understand that changes in the location of the headquarters of large companies have a solid ability to track changes in a country's economy. Because, more often than not, a region has booming industries, a growing population, a bigger market, and a good economy for these local companies to become big companies. On the contrary, when a region's industries are declining, and the economy is depressed, the companies there will also leave the list of big companies.

2.5.2 Headquarters Location reflects Industry Change

There are two paradigms in choosing the location of corporate headquarters. The first assigns a priority to cost minimization, focusing on a city's natural advantages, being geographically central to "multi-locational" markets, and having a robust transportation infrastructure. This paradigm is used by, e.g., raw material, manufacturing, and oil extraction industries. The second paradigm is pursued by companies aiming to enhance productivity. These companies are less sensitive to costs and require well-educated labor to quickly exchange information with other industry cluster members to increase their competitiveness. This is often seen in information technology, health care, and finance industries. Different industries select different locations for their industry clusters. Thus, by observing the changes in the location of headquarters clusters over time, one can gain insight into the industry change.

3 Project Development

This section outlines the key stages of the project's development, covering the history of its progression, the tools utilized, and the data extraction and processing methodologies. Each aspect of the project contributed to the creation of a dynamic web-based visualization of U.S. corporate headquarters locations over time.

3.1 History of Development

The project was developed by a team consisting of Jiacheng Lang and Jonas Greim, who met regularly once a week to discuss the progress, share ideas, and address challenges encountered during the research and implementation process.

In the early stages of the project, the team conducted research to identify existing visualizations or mappings of top U.S. companies over multiple years. However, only a few relevant examples were found, most of which were limited to specific years [Des24].

The next challenge was acquiring relevant company ranking data, specifically focusing on the Fortune 500 and S&P 500 indices. Unfortunately, there was no easily accessible dataset that covers these indices over multiple years. As a result, the team explored various websites to find sources of ranking data that could be scraped. Ultimately, only two websites were identified as offering high-quality, free data. One provides data for the Fortune 500 [Mon24], and the other for the S&P 500 [Fin24]. While Bloomberg provides comprehensive financial data, its cost was prohibitive for this project.

The team used the web-crawling framework Scrapy to successfully scrape data from the two identified websites. Scrapy is a large and powerful library that required time to fully understand and implement. However, the team encountered no significant issues during the process.

A significant problem the team encountered was the lack of headquarters location data in the scraped datasets. To address this, we turned to Wikipedia, which has entries for nearly every major company. Most Wikipedia company pages feature an infobox on the right-hand side that includes the headquarters address, and often provides a clickable link to the exact geographic coordinates.

An attempt was made to automate the extraction of this information with Scrapy. However, inconsistencies in the HTML structure of the infoboxes across different pages made the automation process more complex than anticipated. As a result, it was not possible to scrape the pages effectively.

As the next step, the team attempted to use the official Wikipedia API for data extraction. However, both the API and the Python Wikipedia API wrapper did not provide direct access to the data within the infoboxes. Instead, the API only grants access to the full HTML structure of the page, resulting in the same challenges encountered when scraping the Wikipedia pages directly.

The team also experimented with the new Wikipedia "API:Geosearch" which allows users to send a GET request with a Wikipedia article title and receive the corresponding geographic coordinates [Fou24b]. However, this approach was only successful for a limited number of companies due to the absence of coordinate data for many entries.

One challenge encountered was the significant amount of time required to work with the Wikipedia API. This was largely due to the lack of well-structured official documentation, which failed to provide essential details. Critical information, such as the proper setup of API calls, the full range of API capabilities, and how to formulate accurate search queries, was either unclear or missing entirely, further complicating the process.

The team then searched for better methods of obtaining location data. Based on a recommendation from Dr. Efer, Wikidata was identified as a valuable resource. The key advantage of Wikidata is that all information is structured in key-value pairs, making it possible to access the coordinates through the "headquarters location" key.

However, accessing this information via the API requires knowing the QID (unique identifier) of the Wikidata article. To retrieve the QID of a company, an API text search for the company's article must be conducted. The team performed API title searches using the company names from the Fortune 500 ranking. Unfortunately, this approach encountered challenges. Many company names are either synonyms for common objects, are names of sub-companies, or are ambiguous, often resulting in inaccurate or irrelevant articles being returned as the top result.

To improve accuracy, an API title text search with properties filtering was implemented. The search was configured to ensure that the search term is an instance of P31 (a company) or a subclass of P279* (of a company). This approach successfully retrieved 89 QIDs from 177 company names. The primary reason for this limited accuracy is the absence of essential properties in many Wikidata company articles, which hindered the retrieval process.

To further improve results, an API text search without direct label filtering was at-

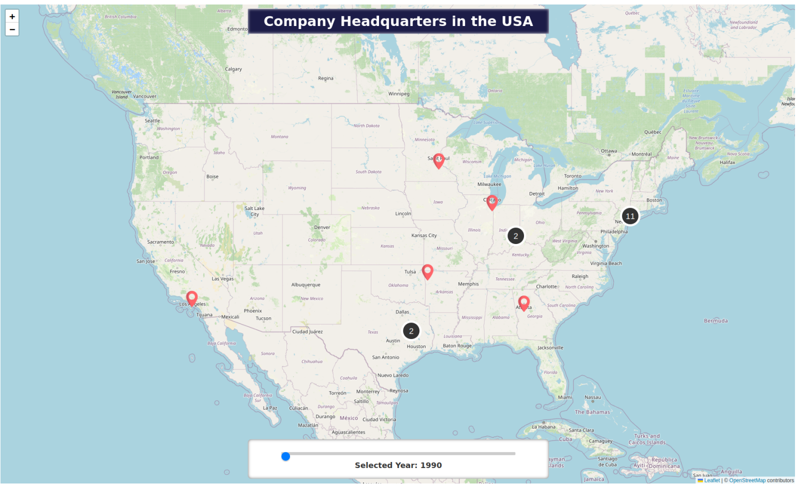

+
−
Company Headquarters in the USA
11
2
2
Selected Year: 1990
Leaflet | © OpenStreetMap contributors

Figure 1: Screenshot of the initial website prototype

tempted, adopting a more trial-and-error approach. This involved searching for the company title and its synonyms (labels) while also checking for the existence of an English Wikipedia article associated with the name. This method significantly increased the hit rate, but also led to a higher false positive rate. The key to improving accuracy was to check for the presence of a key-value pair with the "headquarters location" attribute. If this attribute was found, the entry was highly likely to be the correct Wikidata entry. In cases where Wikidata returned no QID, an incorrect entry, or no location, the search name was manually adjusted. Further details on this approach can be found in Chapter 3.3.

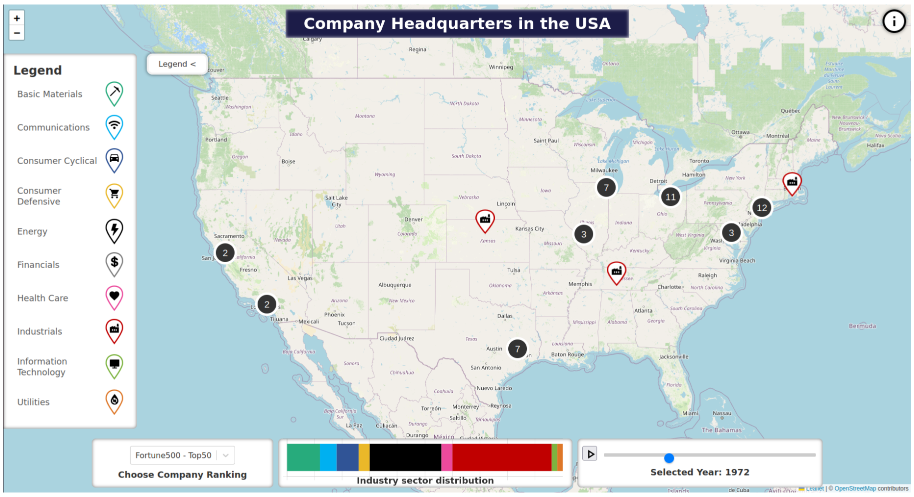
After obtaining a sufficient amount of reliable data, the team decided to explore potential visualization tools, with QGIS and Leaflet.js being considered. Leaflet.js was selected due to its suitability for dynamic, real-time data visualization and its compatibility with web-based applications. Consequently, JavaScript was chosen as the frontend language, with React serving as the framework. React is currently the most widely used frontend framework, and one of the team members had prior experience with it, making it a natural choice for this project.

Following this, work began on frontend visualizations and design, involving multiple design iterations. The first website prototype, developed for the final presentation, utilized only the Fortune 500 data and included a year slider to dynamically visualize headquarters location changes over time. A screenshot of this prototype is shown in Figure 1.

To integrate the S&P 500 dataset, a drop-down menu was added, allowing users to switch between datasets. To visually distinguish industry sectors, we created ten custom icons, each representing a different sector, which were used as map pins for company headquarters. These icons were designed in Figma, with the Iconify plugin serving as the source for the raw icons.

Additionally, to better represent industry changes, a bar chart was developed to display sector distribution for the selected index and year. Another option considered was the use of summarized pin icons with distribution circles to represent industry sectors. However, due to time constraints, only the bar chart was chosen, as it clearly illustrates the full industry distribution for the selected year, providing more comprehensive value.

A feature was also implemented to optionally display the top 55 companies, providing a smoother transition at the ranking threshold. However, since the S&P 500 data source only provides information for the top 20 companies, this feature was ultimately removed.



Fortune500 - Top50

Choose Company Ranking

Selected Year: 1972

Figure 2: Screenshot of the final website application

The frontend and design iterations culminated in the final version of the website, as shown in Figure 2. The website was deployed via GitHub Pages to ensure broad accessibility and can be accessed here [Jon24a]. For a more detailed overview of the website's features and capabilities, refer to Chapter 4.2.

3.2 Tools Used

A range of tools was selected based on their functionality and suitability to meet the project's requirements. Each tool is outlined below, including its role in the project, a brief description, and an analysis of its advantages and disadvantages.

- **Scrapy**
  - **Role:** Data extraction
  - **Description:** Scrapy is an open-source web crawling and scraping framework for Python, designed to efficiently extract structured data from websites. It supports a wide range of use cases, from simple web scraping to complex web crawling tasks [Zyt24].
  - **Advantages:**
    * **Asynchronous Processing:** Handles multiple requests simultaneously, significantly improving speed for scraping tasks.
    * **Built-in Data Extraction:** Supports XPath and CSS selectors for efficient and flexible data extraction.
    * **Middleware Support:** Includes features like user-agent rotation and proxy management (while not needed for this project, it could be useful if different data sources are used).
    * **Item Pipelines:** Enables post-processing of scraped data and supports storage in various formats (e.g., JSON, CSV, databases).
    * **Extensive Documentation:** Scrapy's extensive and well-organized documentation makes it easier to troubleshoot and customize.
  - **Disadvantages:** Scrapy has a steeper learning curve compared to simpler libraries like BeautifulSoup. Additionally, the initial setup can be more complex due to its more sophisticated features.
- **Wikipedia API**
  - **Role:** Data extraction
  - **Description:** The Wikipedia API allows users to retrieve, search, and manipulate content from Wikipedia programmatically [Fou24a].
  - **Advantages:**
    * **Free Access:** Provides free access to the extensive data and content available on Wikipedia.
    * **Direct HTML Access:** Allows retrieval of HTML content without the need for web scraping.
  - **Disadvantages:**
    * **Confusing Documentation:** The API documentation is confusing and poorly structured, making it difficult to understand how to set up API calls correctly or formulate queries.
    * **Limited Structured Data:** The API does not consistently return information in a structured format like key-value pairs. It only provides structured data in a few specific cases, such as citations, images, and page descriptions.

        - **Inconsistent HTML Structure:** The HTML structure of Wikipedia articles is inconsistent, which makes the scraping of information not possible.
        - **Geosearch Limitations:** The Wikipedia Geosearch API is not usable in context of companies, because only a small number of Wikipedia pages include coordinates.
        - **Incomplete Coverage:** Not all companies, particularly older ones, have corresponding Wikipedia entries.

- **Wikidata API**
    - **Role:** Data extraction
    - **Description:** The Wikidata API provides access to an open-source knowledge base that stores structured data. It is part of the Wikimedia Foundation, which also manages Wikipedia. The API utilizes the SPARQL query language for data retrieval [Fou24c].
    - **Advantages:**
        - **Structured Data:** Data is organized into items and properties, allowing for querying and use in automated processes.
        - **Online Query Service:** Offers a web-based interface to test SPARQL queries interactively.
    - **Disadvantages:**
        - **Confusing Documentation:** The documentation is unstructured and lacks clear explanations for setup, query construction, and the property system.
        - **Inaccurate Search API:** Accessing specific information requires the QID of an article, which cannot be retrieved automatically, leading to difficulties in obtaining relevant data.
        - **Incomplete Property Data:** Many relevant properties are missing, limiting the scope and effectiveness of search functionality.
        - **Limited Usefulness of Online Query Builder:** The built-in online query builder does not provide sufficient functionality for effective query construction.
        - **Complex Query Formatting:** Queries must be formatted in specific ways that are not well-documented.
        - **Current Location Data Only:** The API only provides information on current headquarters locations, lacking historical location data.
- **React.js**
    - **Role:** Visualization
    - **Description:** React.js is an open-source JavaScript library developed by Meta, designed for building dynamic and efficient user interfaces through the creation of reusable UI components [Pla24].
    - **Advantages:**
        - **Widely Used:** React is currently one of the most widely used frontend frameworks.

    - **Component-Based Architecture:** Promotes the use of readable, reusable components, leading to modular and maintainable code.
    - **Fast Rendering:** Enhances performance by efficiently updating and rendering components through its virtual DOM.
    - **Good Documentation:** Extensive and well-organized documentation.
  - **Disadvantages:**
    - **Complexity** Challenging due to its ecosystem's variety of libraries, tools, and configuration options.

- **Leaflet.js**
  - **Role:** Visualization
  - **Description:** Leaflet.js is an open-source JavaScript library designed for building interactive maps in web applications. It allows developers to add interactive layers such as markers, popups, polygons, and GeoJSON data to maps, providing a highly customizable mapping solution [Con24a].
  - **Advantages:**
    - **Lightweight and Fast:** Leaflet is designed to be lightweight, making it fast and highly performant, especially for web-based applications.
    - **Responsive Interactions:** The library supports smooth interactions like zooming and dragging, providing a responsive and user-friendly experience.
    - **Overall Experience:** Leaflet successfully fulfilled all project requirements.
  - **Disadvantages:**
    - **Adequate Documentation, but Lacking Context:** While the documentation is generally good, it lacks a clear structure that distinguishes between features, essential functionality, and in-depth function explanations. This can make it challenging to understand how specific features fit into larger applications or workflows.

- **Figma**
  - **Role:** Visualization
  - **Description:** Figma is a web-based design and prototyping tool primarily used for creating user interfaces (UI). In this project, it was used to design and export company sector icons as images [Fig24].
  - **Advantages:**
    - **Real-time Collaboration:** Figma allows multiple users to collaborate on the same design in real-time, directly in the browser.
    - **Comprehensive Design Tools:** Figma offers a wide array of design tools.
  - **Disadvantages:**
    - **Complex Interface:** The interface and toolset can be overwhelming due to the extensive features and options available.

3.3 Data extraction and processing

This section outlines the data sources, as well as the scraping and processing methodologies employed. The code used for these tasks, along with setup instructions, is available in the project's GitHub repository for data scraping and processing [Jon24c].

The company ranking data was scraped from two different websites. For the S&P 500 rankings, data was obtained from *finhacker.cz*, a Czech blog focused on personal finance, investing, and trading [Fin24]. This source provided the top 20 S&P 500 rankings for the years 1958 through 2005.

For the Fortune 500 data, information was scraped from the financial section of CNN's website, *money.cnn.com* [Mon24]. This site offers comprehensive financial news, analysis, and insights, and provided the top 500 Fortune 500 rankings for the years 1990 through 2024.

The data scraping was conducted using the Python framework Scrapy[Zyt24]. For the Fortune 500 data, the scraper dynamically modified the URL to access and retrieve the rankings for each year. In contrast, the complete S&P 500 data was embedded within the HTML of the webpage and was extracted directly. Each set of ranking data was saved as a JSON file.

However, the ranking data did not include information regarding headquarters locations. To address this, the Wikidata API [Fou24c] was utilized to obtain the necessary location data. This extraction and processing workflow was divided into the following steps:

1. Extract all unique company names
2. Retrieve the Wikidata QID for each company name
3. Gather the location and industry data from the corresponding entries
4. Map the unique company data back to the rankings
5. Create a GeoJSON file from the results

The first step involved creating a JSON list containing all unique companies from the rankings. Each entry was enriched with the following attributes: "companyName", "searchQueryCompanyName", "wikiDataName", "qid", "headquarterCoordinates" and "industry".

The "companyName" represents the original name as provided in the data. The "searchQueryCompanyName" is the name used for search queries. The "wikiDataName" corresponds to the retrieved name of the Wikidata entry, while the "qid" serves as the identifier for the Wikidata entry, which is essential for retrieving data from the correct source. The "headquarterCoordinates" and "industry" attributes are used to save the location and indudustry sector data.

The inclusion of these attributes is essential due to the high inaccuracy of the Wikidata entry name searches and the inconsistencies found in the scraped company names. By adding the "searchQueryCompanyName", a mechanism for manual correction is provided. If an incorrect entry is retrieved with an inaccurate "wikiDataName" or missing/incorrect "headquarterCoordinates", it can be easily updated to ensure the accuracy of the data retrieval.

Secondly, the Wikidata Search API was used to identify the corresponding QID for each "searchQueryCompanyName". From the array of returned results, the first QID was extracted and saved to the JSON file.

Subsequently, the Wikidata API QID search was used to retrieve information on each company's headquarters location and industry sectors. After this step, each company was manually categorized into one of ten industry sectors, with assistance from ChatGPT. The results, along with the corresponding Wikidata entry name, were then saved to the JSON file.

Finally, after retrieving all the necessary data, the enriched unique company dataset was mapped back to the original ranking dataset. The JSON data was then converted to GeoJSON format to enable visualization on the frontend.

4 Application

This section outlines the practical aspects of the project's implementation, covering the final tech stack, website features, and the knowledge extracted from the headquarters data.

4.1 Final Tech Stack

The tech stack is divided into two main components: the data scraping and processing segment, detailed in Chapter 3.3, and the visualization segment, discussed in this chapter.

The visualization of the scraped data is implemented through a website [Jon24a]. The source code for this website, along with detailed setup instructions, is available in the associated GitHub repository [Jon24b].

The website tech stack is designed to be as simple as possible. The frontend is built using JavaScript, HTML, and plain CSS. The backend uses Node.js as a static server, responsible for serving the pre-scraped static data in form of two GEOJSON files. There is no dynamic server-side processing or database integration. Additionally, the project employs NPM, the default package manager for Node.js, to manage dependencies efficiently.

To delve deeper, the frontend is developed using React and Leaflet.js, as detailed in section 3.2. React enhances the dynamic and responsive rendering of the map visualizations, allowing users to interact with the displayed data. To facilitate the selection of different years and visualize industry sector distributions, the project incorporates the "react-slider" and "react-chartjs-2" NPM packages [Con24c; Con24b]. Leaflet.js enables efficient map rendering, provides customizable markers, and allows for dynamic interactions with map layers.

The website is hosted on GitHub Pages and is available here [Jon24a]. A GitHub workflow is configured to automatically deploy the build to GitHub Pages whenever changes are pushed to the main branch.

4.2 Website Features and Capabilities

The website is designed as a single-page interface that allows users to interact directly with a map. Users can utilize various interactive tools to explore and modify the displayed data according to their preferences, all without the need for scrolling through the webpage itself. Users can access the deployed website here [Jon24a].

When users first visits the website, they are greeted by a map of the United States, prominently titled "Company Headquarters in the USA." They can easily zoom in and out and drag the map to explore different areas at their convenience.

On the map, icons indicate the headquarters locations of top U.S. companies. Each icon is uniquely colored and designed to represent one of ten distinct industry sectors, making identification easier. The industry sectors are: Basic Materials, Communications, Consumer Cyclical, Consumer Defensive, Energy, Financials, Health Care, Information Technology, and Utilities.

Users can also open a legend on the left-hand side, where the icons are explained. If too many headquarters icons are displayed in a limited space, they are summarized into a single number. Users can click on this number to zoom in and view all the individual headquarters.

Users can also click on a headquarters icon to open a popup that provides more information about the company. This popup displays the company name, a link to the corresponding Wikidata article, the year of the ranking, the ranking placement, and, depending on the selected ranking, either revenue and profit or market capitalization.

At the bottom of the page, users will find three interactive tools. On the left-hand side, there is a dropdown menu that allows users to select the displayed top company rankings, which include the top 20 companies in the S&P 500 index and the top 50 companies in the Fortune 500 index. On the right-hand side, a year slider enables users to select the year for the displayed company rankings. The available year ranges depend on the selected ranking: the Fortune 500 spans from 1958 to 2005, while the S&P 500 covers 1990 to 2024. Additionally, users can click a play button next to slider to gradually cycle through the selected years. In the center, between the two tools, is a bar chart that illustrates the distribution of industry sectors for the selected company ranking and year. The chart uses the same colors as the headquarters icons for easy identification. Users can also hover over the bars to see which sector each bar represents.

In the top right corner the users will find an info icon, that displays a pop up when hovered over. This popup explains that headquarters locations are sourced from Wikidata and, in some cases, only the city, not the exact coordinates are accurate. Additionally, it explains that changes in a company's headquarters location over time are not reflected, as only the most recent location is shown. Unfortunately, Wikidata does not provide historical headquarters location information.

4.3 Extracting Knowledge from Headquarters Data

In this section, we present four tables that display the website data to help you understand geographical location and industry changes. The tables serve as examples, showcasing a portion of the data at 15-year intervals. The website contains more detailed data that you can explore based on your interests and needs [Jon24a].

4.3.1 Location Change

The United States can be divided geographically into four major regions: Northeast, Midwest, South, and West. The data in the two tables is extracted from our Fortune 500 and S&P 500 datasets. The numbers represent the count of headquarters for large companies in each specific region during a a particular year.

Year	Northeast	Midwest	South	West
1958	23	15	8	4
1975	20	16	10	4
1990	19	17	8	5
2005	16	15	12	7

Table 1: Location Change of Fortune 500 Companies by Region

Table 1 shows that the number of headquarters of large companies in the Northeast is decreasing, the Midwest remains stable, and the South and West are increasing.

Year	Northeast	Midwest	South	West
1990	11	5	4	0
2005	8	2	5	5
2020	4	4	5	7

Table 2: Location Change of S&P 500 Dataset by Region

Table 2 shows that the number of headquarters for large companies in the Northeast has continued to decrease, while the Midwest and South have remained relatively stable, with a pronounced increase in the West.

4.3.2 Industry Change

The development of industry changes in the United States can be observed through three stages of economic growth. First, the industrial society transformed into a consumption-driven economy. Subsequently, the high-tech industry emerged, supported by a vast consumer market, financial capital, and venture capital. The industrial sector, as shown in the tables 3 and 4 below, includes manufacturing, energy, and basic materials industries. The consumer sector includes defensive consumption, cyclical consumption, and the financial sector. The high-tech sector includes health care, telecommunications, and information technology industries. The numbers represent the count of headquarters for large companies in each specific category during a particular year.

Year	Industrial Sector	Consumer Sector	High-Tech Sector
1958	37	8	5
1975	37	6	7
1990	32	8	10
2005	12	21	17

Table 3: Industry Change in Fortune 500 Companies Over Time

Year	Industrial Sector	Consumer Sector	High-Tech Sector
1990	6	6	8
2005	4	9	7
2020	1	10	9

Table 4: Industry Change in S&P 500 Companies Over Time

Tables 3 and 4 clearly demonstrate that between 1958 and 2020, the U.S. economy transitioned from being predominantly focused on industrial manufacturing to being dominated by consumption, information technology, medicine, and finance.

5 Conclusion and Discussion

This section outlines the conclusions drawn from our study and discusses potential future improvements in both data collection and application development.

5.1 Conclusion

The trend of large U.S. corporate headquarters relocating away from the northeastern United States continues. In the Fortune 500 dataset, the number of headquarters decreased from 23 in 1958 to 16 in 2005, while the S&P 500 dataset shows a decline from 11 in 1990 to 4 in 2020. The capacity to accommodate headquarters in the Northeastern United States has declined, while the attractiveness of the West has increased significantly.

With the advancement of technology and globalization, the headquarters locations of large companies have become more dispersed. Comparing the data from 1958 to 2005, the number of regions with more than two large company headquarters increased from 7 to 10. Pittsburgh dropped out of the regional list, while the following areas were added: Seattle, Charlotte, Atlanta, and Saint Paul. However, in the context of globalization, high-tech companies have formed the information technology industry cluster in Santa Clara Valley to enhance productivity. Traditional clusters that aimed at reducing costs and improving production efficiency still exist, such as the energy industry cluster in Houston. In summary, although the overall trend in headquarters locations tends to be dispersed, industrial clusters also persist, where company headquarters are concentrated in specific areas to boost productivity or production efficiency.

Through our study, digital humanities workers and social science students can quickly and intuitively obtain information on the geographical location changes, industry shifts, and the rise and fall of different regions regarding the headquarters of large U.S. companies over the decades. Individuals with this macro-image can select an area, industry, factor, or period for in-depth study based on their interests.

5.2 Future Improvements

This section outlines several key areas for future research and development aimed at enhancing the accuracy, scope, and analytical capabilities of the current study. By implementing these improvements, we can gain a deeper understanding of the factors influencing corporate headquarters relocation and industry trends across both regional and global scales.

One of the primary areas for improvement is the precision of the location data for corporate headquarters. As Shilton pointed out [SS99], some major have relocated their headquarters from central metropolitan areas to suburban regions near the city. Currently, the precision of some headquarters' location coordinates is limited to the general geographic coordinates of the city rather than the exact address. In further research, obtaining the precise coordinates of the headquarters' addresses will enable us to more accurately track and study the movement of corporate headquarters from city centers to nearby suburban areas within the same city.

Another limitation lies in the availability of historical headquarters location data. At present, the dataset primarily relies on current location data from the source Wikidata, which do not capture historical headquarters movements. Expanding the dataset to include historical data on corporate relocation would provide a more comprehensive understanding of long-term trends.

Another significant improvement would involve extending the temporal scope of the Fortune 500 dataset, which is currently limited to the terminal year of 2005. By extending the dataset to include data up to 2024, allowing for the analysis of more recent trends in corporate headquarters relocation.

Improving the visualization of industry sectors on the map could be another future enhancement. Currently, icons representing different industries are aggregated when they are located too closely together, which limits the level of detail, particularly in areas with a high density of corporate headquarters. To address this, we propose adding an option that allows users to view all individual industry icons without aggregation. This could be achieved by reducing the size of the icons, making it easier to distinguish individual industries even in dense regions. Alternatively, adopting a method where a circle around aggregated icons visualizes the spread of different industries, as outlined in Section 3.1, could offer a clearer understanding of industry distribution in specific areas.

Another potential improvement involves additional map layers to enrich the analysis of corporate headquarters' locations. By adding historical maps showing population density, natural resource availability, or the distribution of top-tier universities. This additional feature would enable the investigation of the potential dependencies between corporate headquarters' locations and these factors. This approach could reveal whether companies are relocating to optimize access to skilled labor, natural resources, or favorable demographic conditions.

Lastly, to broaden the scope of the analysis, integrating additional ranking indices such as the MSCI World, emerging markets indices, and country-specific rankings like the DAX would enable a comparative analysis of corporate headquarters movements on a global scale.

By implementing these improvements, future research will be better equipped to analyze corporate relocation trends with greater precision and depth. Additionally, it will provide deeper insights into the factors influencing corporate headquarters' location decisions at both regional and global levels.

References

[Mil67] Edwin S Mills. “An aggregative model of resource allocation in a metropolitan area”. In: *The American Economic Review* 57.2 (1967), pp. 197–210.

[Bur77] Leland S Burns. “The location of the headquarters of industrial companies: a comment”. In: *Urban Studies* 14.2 (1977), pp. 211–214.

[SH81] John D Stephens and Brian P Holly. “City system behaviour and corporate influence: the headquarters location of US industrial firms, 1955-75”. In: *Urban Studies* 18.3 (1981), pp. 285–300.

[Ahn84] Leif Ahnström. “Why are offices where they are? The search for factors determining the location of company headquarters”. In: *GeoJournal* 9 (1984), pp. 163–170.

[ARY91] Kasim L Alli, Gabriel G Ramirez, and Kenneth Yung. “Corporate headquarters relocation: Evidence from the capital markets”. In: *Real Estate Economics* 19.4 (1991), pp. 583–600.

[HW91] Steven R Holloway and James O Wheeler. “Corporate headquarters relocation and changes in metropolitan corporate dominance, 1980–1987”. In: *Economic Geography* 67.1 (1991), pp. 54–74.

[GS92] N. Georgantzas and L. Shilton. “Corporate Creativity and Control: Manhattan Office Demand”. In: *International System Dynamics Conference* (1992).

[Por98] Michael E. Porter. “Location, Clusters, and the ”New” Microeconomics of Competition”. In: *Business Economics* 33.1 (1998), pp. 7–13.

[SS99] Leon Shilton and Craig Stanley. “Spatial Patterns of Headquarters”. In: *Journal of Real Estate Research* 17.3 (1999), pp. 341–364. DOI: 10.1080/10835547.1999.12090976.

[Gar+02] Teresa Garcia-Mila et al. “Tax incentives and the city”. In: *Brookings-Wharton Papers on Urban Affairs* (2002), pp. 95–132.

[KT02] Thomas Klier and William Testa. “Location trends of large company headquarters during the 1990s”. In: *Economic perspectives-federal reserve bank of Chicago* 26.2 (2002), pp. 12–26.

[DH08] James C Davis and J Vernon Henderson. “The agglomeration of headquarters”. In: *Regional science and urban economics* 38.5 (2008), pp. 445–460.

[Con24a] Leaflet Contributors. *Leaflet*. 2024. URL: https://leafletjs.com (visited on 10/07/2024).

[Con24b] NPM Contributors. *react-chartjs-2*. 2024. URL: https://www.npmjs.com/package/react-chartjs-2 (visited on 10/07/2024).

[Con24c] NPM Contributors. *react-slider*. 2024. URL: https://www.npmjs.com/package/react-slider (visited on 10/07/2024).

[Des24] Iristorm Design. *American Titans Portfolio*. 2024. URL: https://iristormdesign.com/portfolio/american-titans/ (visited on 10/07/2024).

[Fig24] Inc. Figma. *Figma*. 2024. URL: https://www.figma.com (visited on 10/07/2024).

[Fin24] FinHacker. *Top 20 S&P 500 Companies by Market Cap.* 2024. URL: https://www.finhacker.cz/top-20-sp-500-companies-by-market-cap/ (visited on 10/07/2024).

[Fou24a] Wikimedia Foundation. *MediaWiki API.* 2024. URL: https://www.mediawiki.org/wiki/API:Main_page#Quick_Start (visited on 10/07/2024).

[Fou24b] Wikimedia Foundation. *MediaWiki API: Geosearch.* 2024. URL: https://www.mediawiki.org/wiki/API:Geosearch (visited on 10/07/2024).

[Fou24c] Wikimedia Foundation. *Wikidata REST API.* 2024. URL: https://www.wikidata.org/wiki/Wikidata:REST_API (visited on 10/07/2024).

[Jon24a] Jiacheng Lang Jonas Greim. *Company Headquarters in the USA.* 2024. URL: https://jonasgreim.github.io/leaflet-map-project/ (visited on 10/07/2024).

[Jon24b] Jiacheng Lang Jonas Greim. *leaflet-map-project.* 2024. URL: https://github.com/JonasGreim/leaflet-map-project (visited on 10/07/2024).

[Jon24c] Jiacheng Lang Jonas Greim. *US-headquarter-locations.* 2024. URL: https://github.com/JonasGreim/US-headquarter-locations (visited on 10/07/2024).

[Mon24] CNN Money. *Fortune 500 Archive: Full List 1955.* 2024. URL: https://money.cnn.com/magazines/fortune/fortune500_archive/full/1955/ (visited on 10/07/2024).

[Pla24] Meta Platforms. *React.* 2024. URL: https://reactjs.org (visited on 10/07/2024).

[Zyt24] Scrapy Contributors Zyte. *Scrapy.* 2024. URL: https://scrapy.org (visited on 10/07/2024).