Process book: DataMoviz

Quentin de Longraye, Aymen Gannouni, Victor Le

December 22, 2017

Figure 1: Datamoviz

Contents

You may zoom on all figures using your pdf reader.

1 Presentation of DataMoviz

1.1 Overview

DataMoviz is a web application, that aims at involving the user in a unique experience to discover all the movies from The Movie DB in a very interactive way. This project allows to get impressive insights on the filmmaking scence since the beginnings. What motivates us the most is the opportunity that we have to deliver fascinating observations from raw data through data visualization. There is no restrictions in term of prerequisites for our target audience. Anyone who has interest in movies is heartly welcome to visit DataMoviz!

1.2 Related Work

Who has never watched a movie? Cinema has always been an integral part of many cultures. Even the scientific community had interest in analyzing the movies data. Some works on the Internet Movie Database(IMDb) data have been carried out and can be found in the following publications: [1], [5], and [6]. As data science is gaining more and more interest in the last years, some data visualizations have been also done to uncover insights about IMDb. The following visualizations have inspired us in a certain way to see which potential charts can be considered for the sake of our project: [3] and [2]. Another inspirational visualization from [4] has motivated us to visualize the network of actors.

2 Exploratory analysis

2.1 Data Collection

The used data originates from IMDb. We used queries over the provided API to retrieve the data of each movie using its corresponding IMDb ID. With the collected data, we created a NoSQL database based on MongoDB. This was very helpful, since it allowed to generate very permissive queries. The created database was exposed to an API that was designed using ExpressJS. In fact, we created a dedicated endpoint for each visualization component that required to adapt the data to the filters set by the user. Our backend reacts instantly to the interactions triggered at the frontend by creating a query to retrieve and aggregate the data then returning the results.

2.2 Data Preview

2.2.1 Existing Visualizations

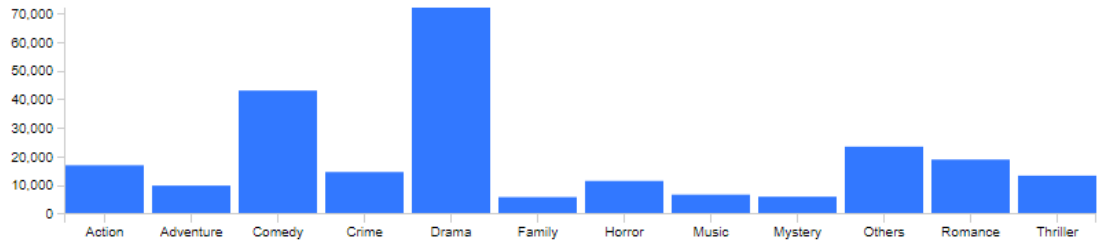
From the many visualizations seen online, the few ones mentioned above in the previous section have particularly inspired us to consider certain charts in our project. The inspirations that we got are listed below:

- From [3]: Use the movie countries and genres as filters for DataMoviz.
- From [2]: Show the evolution of produced countries over the years.
- From [4]: Visualize the network of actors involved in common movies.

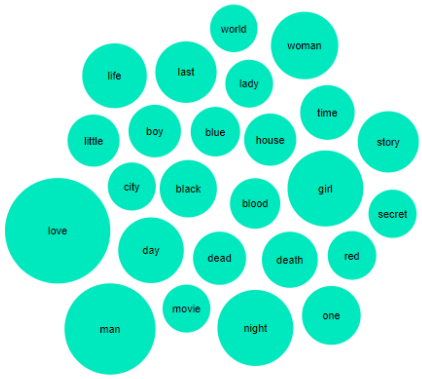
2.2.2 Data Pre-visualization

At the beginning, it was useful to visualize the planned visualization components individually to get an impression of the outcome. In this context, some tools were very handy to start get our hands on the representation of the data. These tools include free online services such as RAWGraphs and Business Intelligence (BI) tools like Tableau. In the following, the charts that have been used at the very first stage of the project are presented:

RAWGraphs While the first chart of the next figure shows the count of movies by genre as a vertical bar chart, the second one represents the single words of the titles as bubbles based on number of occurrences.



(a) Movie count by genre



(b) Movie title keywords

Figure 2: Charts generated using RAWGraphs

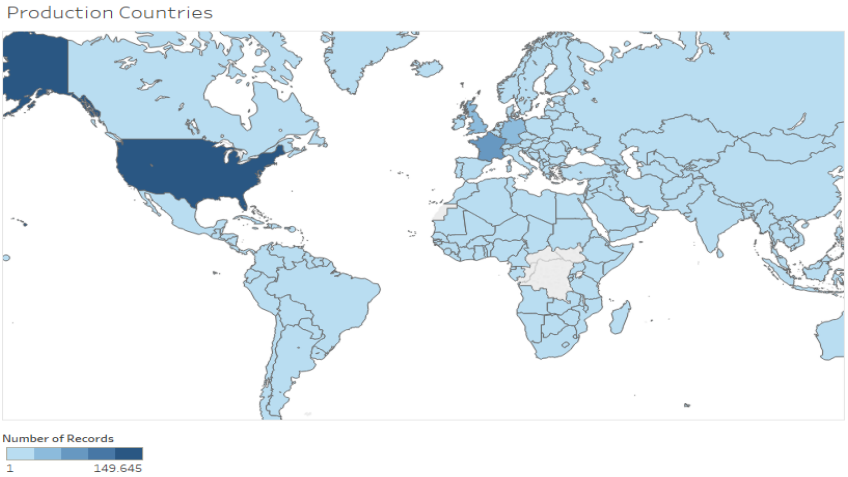
Tableau Tableau is a popular BI tool that is used to create interactive dashboards from input data. Through this tool the charts presented at the top of the next page in figure 3 were evaluated at the beginning.

3 Solution

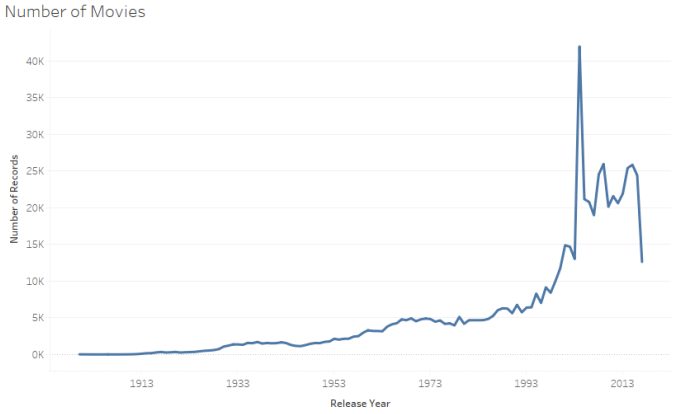
3.1 Considered Visualizations

3.1.1 Initial Ideas

The current implementation of DataMoviz is already very close to the initial proposal. At an early stage of the project, we basically wanted to build an application which allows to see how the production of movies has developed over time. Furthermore, we wanted to find the most influential actors in the movie scene and set a map that displays the major countries that contributed the most to the filmmaking scene.



(a) Choropleth map of production countries



(b) Movie count evolution over time

Figure 3: Charts generated using Tableau

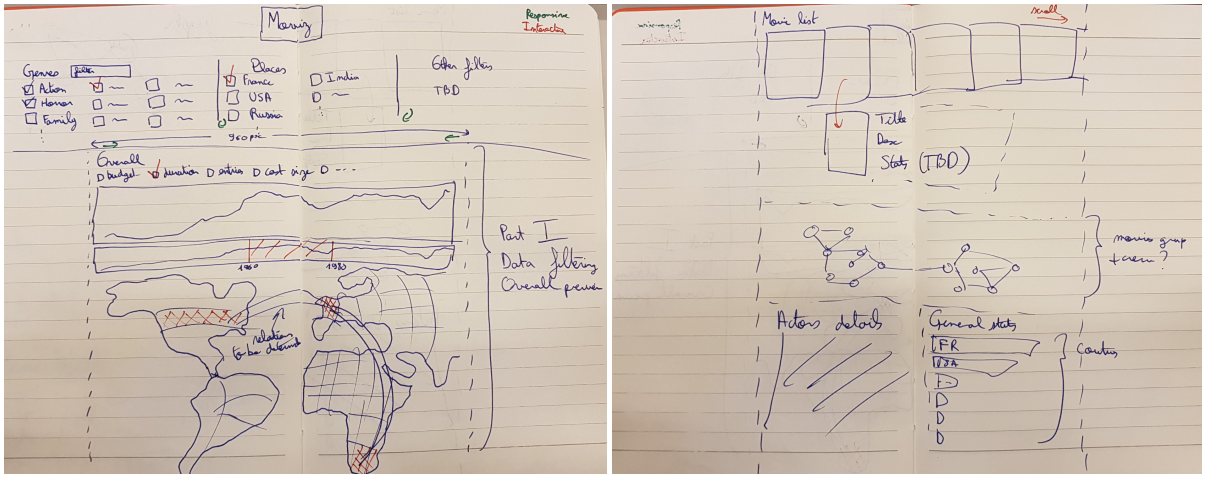
The initial outline of DataMoviz include the following features:

- The evolution of movies over time
- A filter to narrow the time-scale
- A filter for the movie genres
- Details of a certain movie
- The network of actors
- A choropleth map to show the production countries

We will see in the subsequent subsection how we had to deviate a bit from the initial plan to a much better architecture of DataMoviz. Some of the first drafts of DataMoviz can be seen in the Figure 4.

3.1.2 Incremental Improvements

We eventually discovered that the data concealed more insights that had to be revealed through additional visualizations. Therefore after regarding all the possible components that could be included, we decided to add the following elements to DataMoviz:



(a) First page with movies overview

(b) Second page with movie details

Figure 4: A first draft of DataMoviz

- A counter for the number of movies
- A horizontal bar chart for the count of genres
- A donut chart for the proportions of white & black versus color movies
- A bubble chart for the occurrences of words in movie titles
- A slider of posters to ease the recognition of a certain movie

Not only the visualization components needed to be enriched during the project, but also some other crucial changes needed to be done in order to make DataMoviz a filter-oriented application in which the user can adapt the visualization content to his needs.

This allowed for a more natural user-experience, since we did not want to impose any certain walk-through. In fact, we wanted to empower the user will all means to filter over any wished attribute such as time, genre, country etc. Hence Datamoviz is intended to support a wide range of filtering options. In this context, we had to change DataMoviz progressively from a single view to a more modular and decoupled application with a lot of different components that offer the ability to narrow down the analysis of data to smaller subsets.

Given the flexibility of filtering that needed to be achieved, Tableau was used to generate an interactive dashborad with filters set on the movie genres, the production countries and the time scale.

The figure 5 shows the resulted dashboard.

Once convinced of the performance of the filtering concept, DataMoviz was adapted thanks to the created API endpoints.

An other issue that came up in the course of the project was the order of the individual charts and the position of the main filtering components. Thanks to Prof. Benzi advices, we have changed the order of the visualization components and got inspired to deploy a different filtering concept. These improvements that ended to be implemented with full satisfaction are further illustrated in the next section.

3.2 Implementation

3.2.1 Visualization Components

In the section, we present each visualization component and highlights the insights that can be uncovered through it.

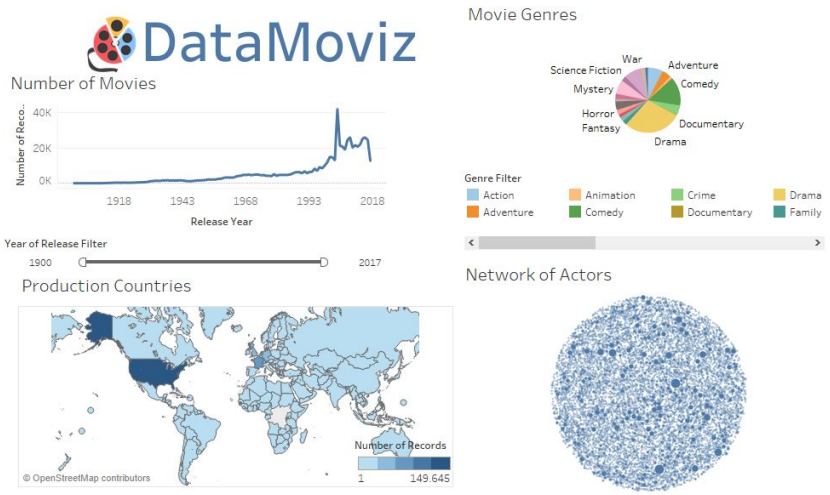


Figure 5: DataMoviz dashboard built with Tableau

The filters bar and panel The filter panel gives the user an overview on the currently applied filters (see fig. 6). It also allows to give an explicit explanation about the current state of the page. In fact, users can remove a filter by clicking on the small cross near the filter name. They also can access the following filters by clicking on *more filters*

- filtering by keywords in movie titles
- filtering by genre of the movie
- filtering by content rating

Figure 6: The filters bar with applied filters

This panel also supports keyboard shortcuts such as *esc* to empty the search fields and close the extended filter section.

The interactive map The interactive map is the first visible visualization when opening the DataMoviz website. This visualization gives a perception of the countries which were involved the most into producing movies with respect to all the applied filters by the other application components (see section 3.2.2 for further details). When the user selects a country, the map colors changes smoothly to give a feedback on the transition between the two states (unfiltered to filtered state). The figure 7 shows what happened after selecting a country (in case of clicking on France).

As we can see in the example above, the intensity of colors decreases for unconcerned countries. It it important to explain the fact that we still have a lot of colored countries, and this because all of them participated in the production of some movies together with French production companies. The color intensity allows the user to even observe which countries were more involved than others in co-production based on current filters. In this case, one can see that Italy and Germany were the most involved countries.

The evolution of movie over time The second visualization shows the evolution of movies along with a time range selector, made with C3.js. It basically allows to see how the number of produced

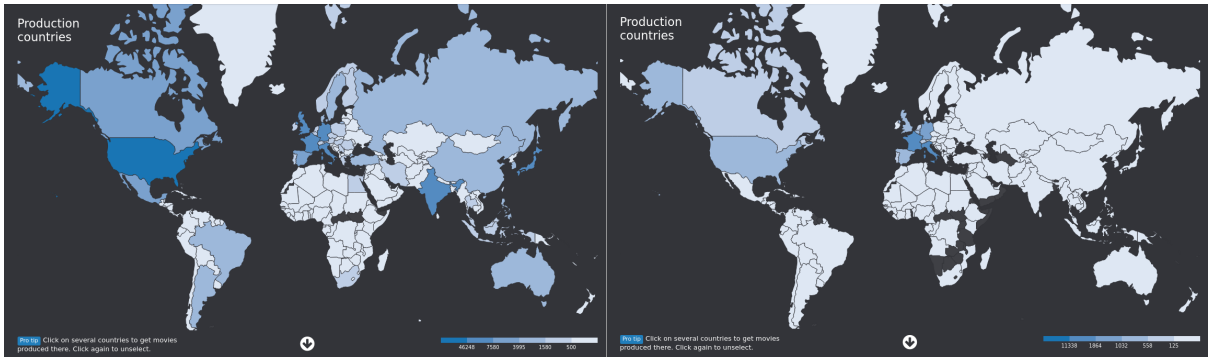


Figure 7: Feedback to country selection

movies developed over time with consideration to the selected filters. Here, users can interact with the visualization by selecting a time range in the subchart (fig. 8). This will create a new filter and restrict all gathered data for a specific time range. All the other visualizations will be consequently updated.

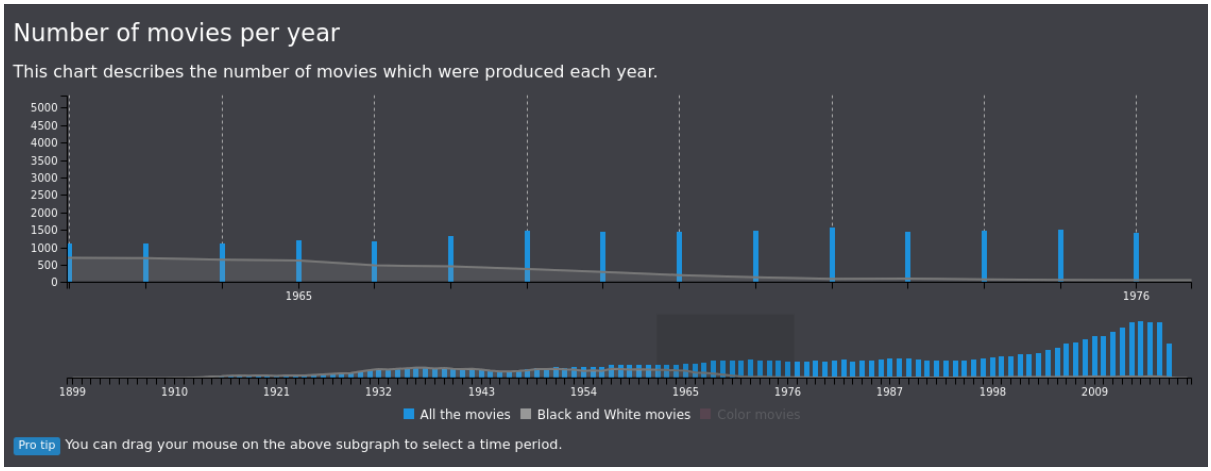


Figure 8: The evolution of movies from 1970 to 1985

In figure 8, we can see that black and white movies were selected (the gray area). Thus, this allows to see the distribution of black and white movies over the selected time range.

The movies statistics DataMoviz displays statistics about the filtered movies. It is possible to see the genres distribution, as well as the color versus black and white movies distribution and the most used words in movie titles (fig. 9). The two first charts were made with C3.js and the last one with D3.js. Some transitions allows users to understand how the distribution evolves when selecting a movie. For instance, in the bubble chart, the size of the bubbles changes dynamically on the screen.

The actors network The actors network is a visualization that shows how the movies are related by common actors (fig. 10a). Each movie and actor are represented as a node in the graph. The color of each node indicates its genre or role respectively. An edge is defined between two movies if and only if there is at least one common actor. Its width is correlated to the number of actors in common. By default, the graph only shows the movies. Nevertheless to show the the cast and crew who were involved, one can just click on the corresponding node. For example in figure 10b, we clicked on the *Shutter Island* node to make its actors visible. The latters are fully connected because they participated in the same movies. There is also an edge from an actor to the movie where he played.

The network is highly interactive and can be customized. We can change the number of actors, crew members and movies to show thanks to some sliders (see figure 11). This visualization offers a way to

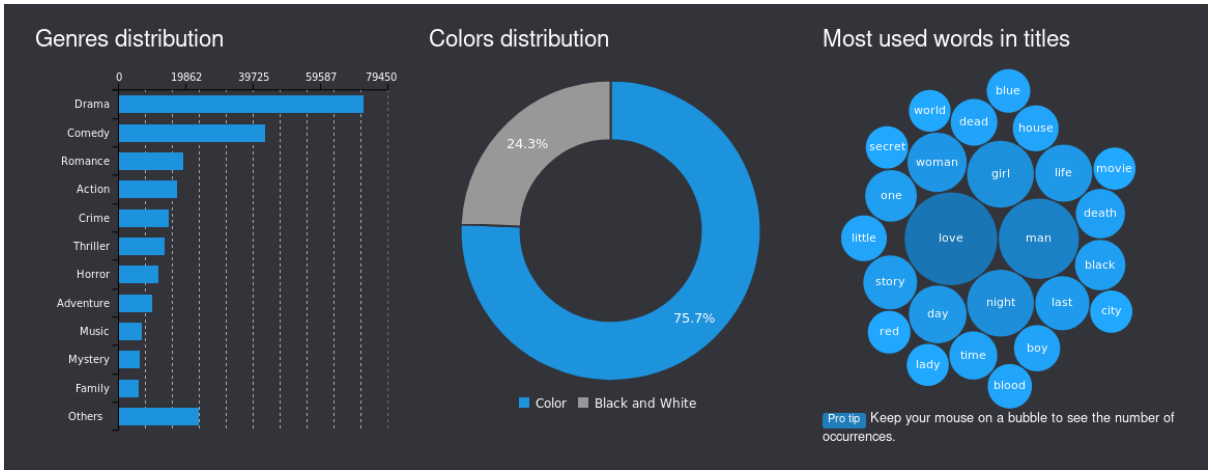
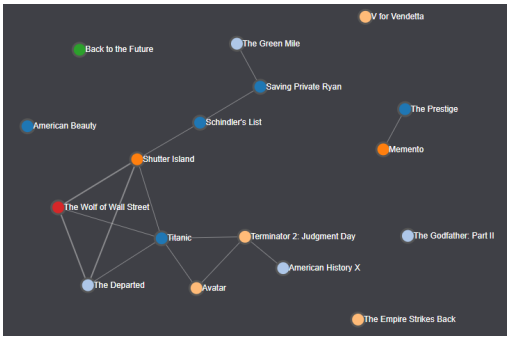
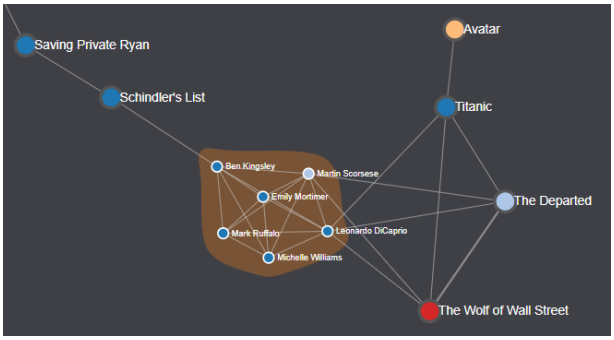


Figure 9: Genres, colors and words distributions



(a) The initial network



(b) The network after clicking on Shutter Island movie

Figure 10: Two views of the network

navigate freely through the network : it supports panning, and zooming. If a user lost the overview of the network, he can still reset the zoom by pressing a button. There is no need to reload the page. Moreover, if one is not interested in having the movie titles or the actor names shown, one can turn it off (Fig. 12). Not to mention that a hover on any node shows this information to indicate the corresponding role or genre.

Figure 11: Sliders allowing to change the quantity of displayed data

The movies details After filtering movies, a list is presented at the end of the page. This offers the user the opportunity to get more information on a specific movie, or to discover new movies matching the selected criteria. The movie details (fig. 13) presents several metrics including the total budget and revenue, its rating and popularity, and a short description. Depending on the movie rating, the color of the plot may turn yellow or red if the rate is too low.

3.2.2 Filtering Concept

The filtering system is one of the most important features of DataMoviz. It allows to interact with the page and ensure that all graphs are in the same, coherent state. Without this concept, we would not have been able to provide such a visual-appealing interactive web application. It was built by creating a

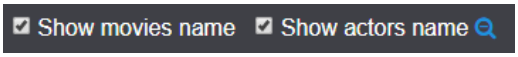


Figure 12: Checkboxes allowing to tweak the display

Figure 13: The movie details for V for Vendetta

single shared state between all components. In fact, each time a component needs to update the filters, it triggers an event which is related to the other components. The following draft (fig. 14) describes how the system works. It was designed before the implementation.

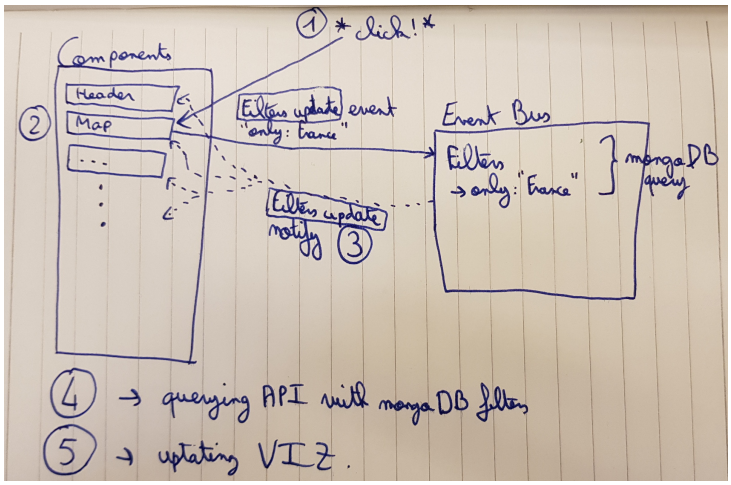


Figure 14: Initial draft of the filtering system

When the user interacts with the page (for instance by clicking on a country of the map visualization (1)), a filter update event is triggered into an event bus (2). This event contains all the filters that are currently applied, with a small addition or deletion (the example shows a constraint in which the user wants to apply a filter of only movies produced in France). The event bus redirects this filter to all the subscribed components (3). Based on the newly set filters, they may query the API (4) to get refreshed data, and update the visualization consequently (5).

This system was designed thoughtfully by following simple computational thinking concepts and by reducing the filtering complexity to a single, atomic and context-independent (stateless) set of filters which can be directly passed to MongoDB. We also take advantage of HTTP cache to allow the browser to cache all previous requests, so we potentially avoid overwhelming processing on the server.

The final implementation is presented in the figure 15. This event happened after clicking on France. We can see in the event payload that an object is stored. This object describes a MongoDB filtering subquery which may be used by the server to restrict the data that should be used during data selection, aggregation or counting.

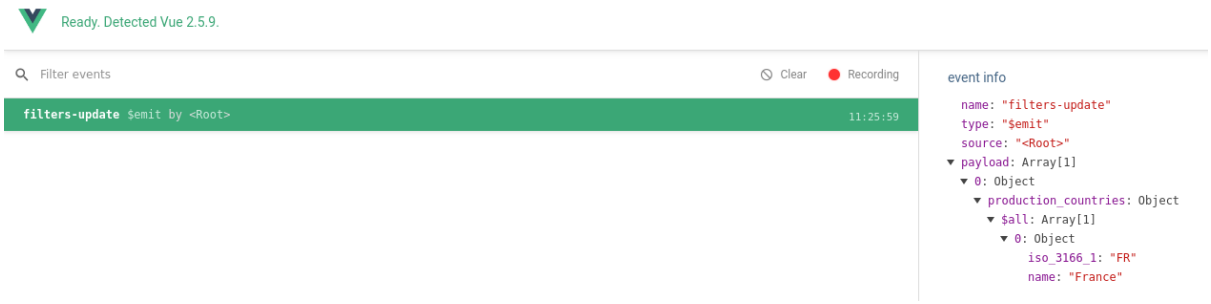


Figure 15: An emitted event when selecting a country

This system is extremely flexible as it allows to add new components without impacting other parts of the application. Each new component should only define two behaviors (if required): filters update triggering after user interaction, and filters update handling when another component updates the filters.

3.3 Evaluation

We discovered a lot of insights about movies when working on the visualization. The first interesting fact was about the most used words in movie titles. It was amusing to see which words are more used in titles, and how it evolved over time. We also discovered how each decade of the twentieth century influenced the movie genres, such as westerns during the 40's or porn movies during the 70's. Our visualization also brought some interesting information on outliers, such as the four movies from Greenland.

Our questions were to better understand the distribution of movies distribution and how the movie scene has developed over time. Thanks to the implemented filters, we were able to understand how countries were involved in producing movies over time and by comparing time ranges. The actors network was very handy to understand some unexpected relationships between movies through its cast.

The final visualization works as expected. Some further improvements could be done to enhance the overall performances, by implementing a requests aggretator in charge to delay and aggregate the requests to the server into a unique query. This would reduce the data traffic between the page and the server, and would certainly boost the overall loading performance.

The application was built using mobile-first methodology which allowed us to bring a fully responsive and working application on mobile phones. Expected performances are however high and some smartphones may have some difficulties to render the full page. This may be improved by limiting the number of queries to the server as stated previously, or by limiting the SVG usage (which would require some major changes). Finally, some minor improvements can be done to enhance the compatibility with other browsers. From a visualization point of view, more visualizations could be added as well as more features such as movies suggestions, random filter selection or more interactions with the charts.

4 Peer assessment

4.1 Quentin de Longraye

Preparation Everybody was prepared for meetings which happened during exercises sessions.

Contribution Everybody contributed to the elaboration of the application, mainly based on skills and interests. Ideas came from each of us to improve current the application gradually, fixing bugs and improve display.

Respect of others' ideas Almost all ideas were implemented and tested to determine their relevance. We tried to promote discussion in the group to understand which parts of the application should be improved.

Flexibility In case of disagreements, we compromised most of the time, or asked for an external point of view.

4.2 Aymen Gannouni

Preparation The lab sessions have helped us a lot to get prepared for the project and know the challenges that may face us. It was always easy to plan the tasks within the group.

Contribution Everyone contributed greatly for the success of the project especially the internal task assignment that respects everyone skills and interests.

Respect of others' ideas Discussions in the team were very constructive and helped for many improvements along the project

Flexibility The working atmosphere was very cool and we could easily coordinate between us without long debates.

4.3 Victor Le

Preparation The preparation was enough.

Contribution Yes, they contributed to the realization of the project and added their stone to the building.

Respect of others' ideas Friendly group atmosphere.

Flexibility No major disagreements.

List of Figures

References

[1] Klaus Dodds. "Popular geopolitics and audience dispositions: James Bond and the internet movie database (IMDb)". In: *Transactions of the Institute of British Geographers* 31.2 (2006), pp. 116–130.

[2] *IMDB Data Visualizations with D3 + Dimple.js - Andrey Kurenkov's Web World*. `http://www.andreykurenkov.com/writing/visualizing-imdb-data-with-d3/`. [Online; accessed 22-December-2017].

[3] *IMDB Movies Visualized*. `https://public.tableau.com/en-us/s/gallery/imdb-movies-visualized`. [Online; accessed 22-December-2017].

[4] *Interet Movie Database(IMDb) Visualization*. `http://gigapan.com/gigapans/4306`. [Online; accessed 22-December-2017].

[5] Andrei Oghina et al. "Predicting imdb movie ratings using social media". In: *European Conference on Information Retrieval*. Springer. 2012, pp. 503–507.

[6] Verónika Peralta. *Extraction and integration of movielens and imdb data*. Tech. rep. Tech. rep., Technical Report, Laboratoire PRiSM, Université de Versailles, France, 2007.