

# **Le Monde**

Visualisation of article contents over time

Hugo Bouigeon

Dec. 2022

# Table of contents

Introduction . . . . .	1
<b>1 Dataset</b>	<b>2</b>
<b>2 Visualisation workings</b>	<b>3</b>
2.1 Word and date input . . . . .	3
2.2 Day selection . . . . .	3
<b>3 Design choices</b>	<b>5</b>
3.1 Graph choices . . . . .	5
3.2 Interactions . . . . .	5
Conclusion . . . . .	6
Annexe . . . . .	7

## Introduction

Events, whether they are about natural catastrophes, geo-political conflicts or scientific discoveries, always bring consequences with them and are themselves often consequences of past events. Newspapers are one of many interfaces where we can find information about the world, and their archives contain valuable historic information. Among them, Le Monde is one of France's more prominent newspapers, which archives articles since 1945.

My goal with this visualization is to give a simplified view over what happened over time in the world, using information contained in articles of Le Monde. For example, one should be able to see that many articles talk about Football in December. Moreover, this visualisation also has the purpose to act as an intuitive interface to navigate faster and smarter between the archives of Le Monde.

This report will provide an overview over how the data to be visualized was collected. It will explain how the visualisation works and describe design choices.

# Chapter 1

## Dataset

As I did not find an adequate dataset online for what I wanted to do, I wrote up a scrapper to collect data directly from the archives from Le Monde on <https://www.lemonde.fr/archives-du-monde/>. The python-script itself worked by opening and downloading the HTML code of each page of the archive site, which contains titles and other information about up to 40 articles. By analysing the HTML code I could then retrieve the information I wanted. Normally one would use an API to do this, but I did not find any freely available ones.

Using this script, I created a CSV file with columns containing the following information:

1. Id for the date of an article
2. Date of article (in YYYY-MM-DD format)
3. Hyperlink to article
4. Title of article
5. Authors of the article
6. Description of the article
7. Address of the teaser-image for the article
8. Category of the article ("international", "politique", "sport", etc.)

I run the script to save data for all articles from the 15th of December 2021 to the 15th of December 2022, which corresponds to roughly 35,000 articles.

## Chapter 2

# Visualisation workings

The goal of the visualisation is to show how often a certain word occurs in the articles published during one day, over a given time span. For instance, the default settings when one opens the page are "Monde" being prompted in the search bar, with data being visualized for the year of 2022.

### 2.1 Word and date input

If a word is prompted into the search bar, the histogram's rectangles update their height in accordance to the number of times the word was found in the day's titles and descriptions. Words are sought in the same manner as a ctrl-F search does. For example, if the input is "organ" and an article contains the word "organise", this will be counted as an occurrence, which can be troublesome in some cases. A half-violin plot on top of the histogram allows for a smooth representation of the density of occurrences of the input word.

As an additional feature, the prompt "\*" allows to count the number of articles.

The word input is also used to highlight images of articles that contain said word (see Section 2.2). The word input can also be done by selecting the top trending words of the day.

It is also possible to give two different dates to select the time-span one wants to analyse. The histogram and violin plot are then scaled according to the bounding dates.

### 2.2 Day selection

It is possible to click on the bars of the histogram to select a given day. By doing so, several actions are triggered:

1. The date appears in the center.
2. A grid of images appear, which are teaser-images for that given day. Hovering over the image will enlarge it slightly and makes the article name appear on screen (though without uppercase letters, accents, dots, etc.) It is possible to click on an image to open the article itself in a new tab.
3. A horizontal stacked bar chart appears that shows the proportion of articles for each category (international, politique, etc.). Hovering over the category makes its title appear. You can select the category on the chart to highlight article images that correspond to that category.
4. The 12 words which appear most often in this days article descriptions and titles are displayed (common words from the French language are filtered out). They are sorted

by occurrence, and the most common ones are more opaque than the less common ones. By selecting one of these words, it is prompted as a new word input and the graph gets updated.

Articles which are shown are chosen randomly among the days articles, and not all are visible at once. By clicking on the "Le Monde" logo in the center, it is possible to re-sample the images to be shown, and view another random batch of them.

# Chapter 3

## Design choices

### 3.1 Graph choices

The main focus of the visualisation is the histogram which is doubled by a density plot. I believe the density plot is a more suitable way to present the data overall, since it can better highlight when a certain trend was important, even if it was for a single day, whereas a single day on the histogram can be a very thin rectangle. On the other hand though, the histogram gives more precision to the visualisation and most importantly provides an intuitive way of selecting a day in it.

I chose a bar chart to visualise category proportions instead of the more classical pie chart, mainly because it fitted better on the page. Initially Le Monde provides more categories than are shown on the chart. I didn't want to have minuscule categories, nor did I want a huge Other category. Hence I grouped some categories like "Sport" and "Football" or "International" and "Afrique" together and put remaining categories into the "Autre" section.

At first I thought of creating a word-cloud out of the trending words, but putting them into two columns was more in line with the general very symmetrical design of the page, and allowed for word rankings to appear more prominently.

### 3.2 Interactions

I wanted the design style to be very simplistic and intuitive to use, to follow the Le Monde's quite sober design. Colors are pastel and elements well organized to give the whole a symmetrical look.

Clickable content is always highlighted in a way to add to the visualisations intuitivity. As a default, objects opacity is altered, except for the teaser-images and trending words, whose opacities are already meant to highlight selected articles or the trend's ranking. The histogram's bars are also a bit widened to prevent selection errors when the mouse moves slightly. And if the mouse stays in place long enough, a small indicator shows the date and number of occurrences of the input for the day.

## Conclusion

d3 gave me the tools to create this interactive visualisation, which allows for an ergonomic way to navigate easily through Le Monde's Archives by showing when what trending words are most common in articles. Through its mechanics, it provides an easy and smart way to correlate research prompts and a timeline.

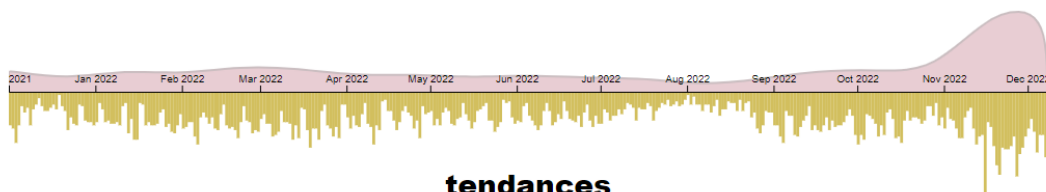
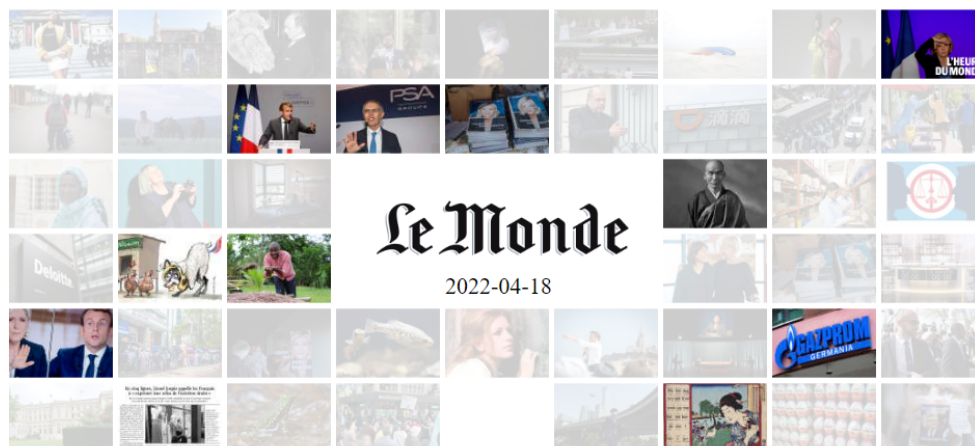
Additional information like view counts on articles could have been useful to give more weight to some articles than others, and to give a possibility to view these articles. Unfortunately I could not retrieve this information with my scrapper. Further improvements could have been to get correct text with accents, uppercase, etc. and maybe to add information authors which I didn't end up using.

Through my time tweaking the code, it actually made me remember some events of the past year such as the presidential elections or the pandemic which apparently was still very present in early 2022. Hence I believe the main purpose of this visualisation was achieved.



## Annex

Input:  Time span:  to



presidentielle	guerre
marine	ukraine
pen	pays
2022	l'election
macron	droite
emmanuel	programme

Page when a day is selected.