Project Proposal: Journalist Deaths Across the World

Karen Sanchez Trejo
Indiana University Bloomington
School of Informatics and Computing
karsanc@iu.edu
(residential section student)

Arnav Arnav
Indiana University Bloomington
School of Informatics and Computing
aarnav@iu.edu
(online section student)

Abstract

Journalism is one of the most important occupation in a free world. Hard working journalists spend enormous amounts of their time and energy to dig into the situations and provide us with information that most people would miss otherwise. Good journalism keeps the organizations in power in check and empower the citizens in a democracy with knowledge about the world. The job however comes with its risks. Over the years journalists have had to face various perilous situations and many of them have been fatal. In this project, we aim to visualize the information about the journalists who gave their life for the job and attempt to understand any reasons that gravity of the situation and aim find out which topics and geographical regions have proven to be more fatal than others.

Keywords : Journalist deaths, Time series, Maps, Network graphs

Introduction

One of the most noticeable attacks on the freedom of press in the recent past has been the attack on the office of the French satirical newspaper Charlie Hebdo in Paris in 2015. In the tragic event two armed men forced their way into the office and went on to shoot journalists in the head (Wikipedia contributors 2018a). The reporters in the office have been vocal in asking for protection ever since (Marlowe 2018). The newspaper employs a private security firm for their offices but reporters believe that spending huge amounts on security of press in a democratic country should not be necessary in the first place.

More recently the reports from Reporters Without Borders about the two journalists who are missing since 2017 emphasize the need for protection of freedom of speech. Samar Abbas, a Pakistani blogger and the founder of Civil Progressive Alliance Pakistan, a group that protects human rights and religious freedom was abducted along with four other bloggers and remains missing. The Bangladeshi reporter Utpal Das, writing for a news website, went missing when he was preparing a story about the ruling party. Authorities have found no information on either of the reporters thus far(Reporters Without Borders 2018c). The International Federation of Journalists report that lists journalists killed in 2018 across the world, in both targeted killings and accidents, supports the argument that journalists should be protected (International Federation of Journalists 2018b).

There have been numerous such attacks on journalists in various parts of the world, and these continue to happen. The Committee to Protect Journalists (CPJ) is an American non-profit organization that was founded in 1981 in response to the harassment of the Paraguayan journalist Alcibiades Gonzlez Delvalle. The committee has been advocating press freedom ever since and holds gatherings to honor the journalists who face attacks, intimidation and even prison in order to report the world events. The CPJ maintains a running list of the journalists killed across the globe, and publishes a magazine and surveys that detail the accounts of such attacks and other information about the freedom of the press (Wikipedia contributors 2018b) (Committee to Protect Journalists 2018c).

Other important organizations working to protect journalists include the Reporters Without Borders, and the Council of Europe. The Reporters without Borders (RSF) is a non profit organization based in Paris, advocating freedom of press and information. The RSF has been actively, since its foundation in 1981 and to protect journalists from attacks, reduce censorship of the press and aid persecuted journalists and provides for the safety of war correspondents (Wikipedia contributors 2018e). The RSF has been campaigning for the creation of a United Nations body under the Secretary General for the protection of Journalists (Reporters Without Borders 2018d).

The Council of Europe is an international organization that aims to uphold human rights, democracy and the rule of law in Europe (Wikipedia contributors 2018c). The council has been working on the freedom of press and dissemination of information and maintains a platform that compiles threats and concerns about the freedom of media in order to better address the attacks against media professionals (Council of Europe 2018).

The International Federation of Journalists is the largest global union of journalist trade unions in the world. The federation first founded in 1926 is aimed at working for social justice, democracy and human rights among other areas, now has 600,000 members in more than a 100 countries (International Federation of Journalists 2018a).

In this project, we aim to visualize the deaths of journalists across the world from 1992 to March 2016, and hope

to shine some light on these deaths, provided by the CPJ on Kaggle (Committee to Protect Journalists 2016). A good visualization would help understand the importance of the issue and aid in further analysis of the problem to and help minimize the dangers faced by media professionals and help freedom of press.

Background and Related Work

Protection of media personnel and staff and freedom of the speech and press are important issues and various important works have preceded this one. Here we list some important works and visualizations related to the same topic as this project.

A report (Reporters Without Borders 2018b) published by the RSF in September 2018, anticipates an increase in violence against reporters and journalists in Germany in 2018 and urges the authorities to take appropriate measures. There have already been various threats and attacks on journalists in Germany following the anti-immigrant protests, and counter protests. The RSF does not however have any concrete data or statistics estimates yet, that support this anticipated increase in violence.

A Google News Initiative report (Whitaker 2017), in an outreach program blog, outlines the efforts made by Google to push forward cybersecurity initiatives for online news websites and protect online journalists from cyber attacks, such as Distributed Denial of Service (DDoS). The article describes the severe lack of security measures taken by the news websites and proposes measures to mitigate possible cyber-attacks.

An article from July 2018 by The Hindu news (The Hindu 2018), indicates that the violence towards journalist in India has increased. Reporters without borders said that in 2017 four journalist were killed and another four during the first six months of 2018. A media watch group demands action from the government, who has not been acting as it should. They also ask to condemn the online campaigns of hate and harassment towards journalists. The attacks aim to cause fear jeopardizing the freedom of the media.

Live Mint, an online news portal, in a 2017 article (Kishore 2017) describes the re-ignition of the debate for the protection of journalists following the death of the journalist Gauri Lankesh. India ranks poorly in the list of high impunity countries, and many of the journalist deaths remain unsolved. It was further reported that Indian journalists have been dedicated activists for the Right to Information (RTI) initiative, and regularly face prejudice from officials. The article calls for legal action to make the country safer for journalists.

Balguy-Gallois in (Balguy-Gallois) calls attention to the increase in journalist deaths in war torn areas. The author proposes that these journalists should be protected as citizens and reiterates the humanitarian law which makes any attacks on journalists and media personnel and media equipment illegal, unless the medial is used to instigate breaches to the humanitarian law. The author addresses the need for new equipment for the protection of journalists and suggests the importance of precautionary methods in such situations.

A 2014 article by the Columbia Journal Review (Langley 2014) addresses violations of free expression across Europe with the help of a data map of the continent in Figure 6. The article emphasizes the importance of such visualizations in bringing these attacks into common knowledge and highlights the seriousness of such attacks. The visualization here can be confusing for various reasons. Firstly, the dots on the map are too big, and countries with multiple dots, (specially eastern European regions in the map) can be hard to distinguish from each other. This can be improved by using a color suitable color map and representing the data on the map as a heat map or a density map, with the help of colors instead of dots.

The Reporters Without Borders in figure 3 visualize the freedom index of different countries in the world, on a world map. The countries are ranked based on their freedom scores and users can learn more about the specific countries. The information about the countries includes the general situation and conditions of the media and the press and lists recent relevant news articles. The visualization can be improved by adding other information about other variables such as economic conditions in the countries, and it would help in comparison to provide context of freedom indices of the countries in the previous and following years.

The CPJ website (Committee to Protect Journalists 2018c) displays visualizations of the data set that in this project we aim to analyze. It includes a data map in figure 1. The counts of deaths are represented with circles, which area is proportional to the count. When we represent data with area the viewer could be mislead, people underestimate larger quantities because they do not look large enough. Another problem of this visualization is that in some cases it is hard to distinguish to what country each circle belongs. Although, this problem is solved adding labels that appear when one clicks on the circle. We can also find a histogram that shows the number of deaths by year in figure 2. In this project, we will analyze this variable as a time series which usually is better to visualize as a line chart instead of an histogram. In all visualizations display in the CPJ website there are filters so the viewer can select what data they want to see.

The dataset that this project aims to analyze and visualize was obtained from Kaggle (Committee to Protect Journalists 2016). In this website, users can upload the work that they have done on the dataset. In (Lara 2016) the user displays the data using the programming language R. They show the total deaths by year as a time series. We will not only visualize the data but also do exploratory data analysis, finding a trend and possible explanation of the variation. The user also displays some of the categorical variables histograms that show the count of deaths. We can find some mistakes in this visualizations, like they used colors without having any meaning.

In another Kaggle kernel (Adhokshaja 2015), the user used the programming language R. They display a time series by each year, since the dataset goes from 1992 to 2016, there are too many lines and the plot cannot be interpreted. A better way to do this would be displaying monthly subseries. A common mistake that we can find in this work is that they plot a lot of categories distinguishing them by col-

ors. In this case it is better to group the categories and only display a few colors, and use other graphical cues to distinguish between different groups.

Seelke in (Seelke 2018), in a Congressional Research memorandum, addresses the issue of violence and threats against journalists and media personnel in Mexico and performs detailed comparisons with other Latin American countries. The work includes two important visualizations. Figure 4 compares homicides of Journalists and media workers in Mexico as reported in three different databases from the year 2007 to 2017 using a bar graph. Figure 5 compares the difference perpetrators of violence against journalists in 2017 in Mexico using a well labeled pie chart in 2017. In the first visualization in figure 4, although we see a trend, it might be difficult to interpret since it is a bar chart. A better way would be to augment the bar chart with a a line chart comparing the trend obtained from the three datasets. The pie chart in figure 5, is not a good representation of quantities, since using areas (or sectors) to represent a scalar quantities can be misleading. A better visualization would be to augment the chart with a bar graph comparing those values.

The National Coalition Against Censorship (National Coalition Against Censorship 2018) uses labelled data maps in figure 7, and chronological explanations in figure 8 to visualize censorship in books and music in the United States up until the late 2000s. These visualizations are better for website, or digital media as they include quite a bit of text which can be difficult to read in print unless the font is big enough.

The works stated in this section establish a strong foundation for the project. In this project we aim to provide helpful visualization to help understand the problem better and help further the cause.

Techniques

- **Data Maps:** Geographical Information Systems (GIS) are applications that capture analyze and visualize data that is linked to geographical locations. Mapping data to geographical locations properly can help users to generate interactive queries, and effectively analyze spatial information (wiki.gis contributors 2018a). It is helpful in such applications to visualize the data on a geographical map on a particular location of the world. Since the data we have from CPJ (Committee to Protect Journalists 2016) has countries of the deaths of journalists, we can visualize the data with data maps. If the data has multiple variables associated with each record in the data, it may be helpful in some cases to generate overlays on the geographical maps. Overlaying is a technique in GIS where we combine different variables in the data into one. This helps us better understand patterns in the data such and perform operations such as intersections and subtractions if any meaningful patterns emerge (wiki.gis contributors 2018b).

  Various python libraries are available that make working with GIS data easier. GeoPandas (geopandas 2018) is an extension of the Pandas library in python that allows using Pandas like operations for GIS data. it uses shapely to perform geometrical operations and matplotlib to visualize the data. Henrikki Tenkanen in (Tenkanen 2017) lists various libraries and describes in detail how to use of python for automating GIS process.

- **Network Graphs:** Representing data as graphs can allow us to visualize and discover relationships that might be missed otherwise. A graph is defined in graph theory as a set of vertices and edges. each vertex represents a data point and connections or relations between the vertices are represented by edges(**?**). Representing data as graphs allows us to use numerous sophisticated mathematical techniques that have been developed over the years, to perform a deep analysis the data.

  Networkx (Networkx Contributors 2018) is a python library that allows creation, manipulation and visualization on complex graphs and networks. The visualization library Plotly (Plotly contributors 2018) now also supports network graph visualizations, and can visualize networks created using Networkx.

- **Time Series:** A time series is define a quantitative variable through time. The samples should be taken at successive equally spaced points in time. The usual exploratory analysis for a time series is to fit a trend and then analyze the variation that is left in the residuals (Wikipedia contributors 2018f). Trying to identify possible causes of the oscillations in the residuals after fitting a trend can help us understand and explain the patterns in the data (Wikipedia contributors 2018f).

- **PCA:** Principal Component Analysis is an orthogonal transformation that converts the variables into a set of linear combinations called principal components. The first principal components account for most of the variation in the data (Wikipedia contributors 2018d). It is commonly used for data reduction, but it also allows to get insights from the data. Scikit-learn (scikit-learn contributors 2018) package in python contains an optimized implementation of PCA.

- **Factor analysis:** The purpose of factor analysis is to describe, the covariance relationships among many variables in terms of a few underlying, but unobservable, variables called factors (UCLA Image Institute for Digital Research and Education 2018). In Python there is a package called FactorAnalizer (Biggs 2018) that performs factor analysis. It has the option to rotate the results to get a better interpretation. The estimation can be done using minimum residual or maximum likelihood estimation. The output are the loading of each factor.

Dataset

The dataset provided by the CPJ on Kaggle (Committee to Protect Journalists 2016) contains information of journalist deaths from 1992 to March 2016 in 105 countries. Each event has information about the journalist's work and the murder. CPJ includes in its definition of journalists the staff, freelancers, bloggers, stringers and journalists citizen who report news in the public domain regardless of the medium, which could be printed, photographic, radio, television or

electronic. The dataset includes 18 variables which are described below:

- *Motive:* Categorical variable that describes the motive of the murder. It has three categories: motive confirmed, motivate unconfirmed, media workers. The variable media workers shows if the journalist was a key part of the investigation.
- *Date:* Date of the murder. It contains year, month and in some cases day.
- *Name:* Journalist name.
- *Sex:* Categorical variable that has two categories: Female, male.
- *Country killed:* Country where the assassination occurred. The dataset contains 105 different countries.
- *Organization:* A categorical variable representing the organization where the journalist worked.
- *Nationality:* A categorical variable with 170 labels that describes the journalist nationality.
- *Medium:* A categorical variable that describes the medium where the journalist collaborated. It has 15 labels with the combination of radio, television, print and internet.
- *Job:* A categorical variable that describes the jobs performed by the journalist. It is a combination of internet reporter, print reporter, broadcast, camera operator, editor, photographer, producer, publisher, technician and columnist.
- *Coverage:* A categorical variable that describes the general topic the journalist worked on. It includes a combination of crime, politics, war, corruption, culture, human rights, business and sports.
- *Freelance:* A binary variable that describes if the journalist worked as freelance or not.
- *Local / Foreign:* A binary variable that refers to the origin of the journalist. It has two labels: local and foreign.
- *Source of fire:* A categorical variable that describes people or entities that probably were responsible of the assassination. It includes: criminal group, government officials, local residents, military officials, mob violence, paramilitary group, political group, unknown fire.
- *Type of death:* A categorical variable that contains the classification of the murder only in the case where the motive was confirmed. The labels include crossfire/combat-related, dangerous assignment, murder, unknown.
- *Impunity* (for Murder): A categorical variable. Monitoring conducted by CPJ about the legal process for confirmed murder cases. It has three labels: yes, no and partial. The partial level refers to situations when only some of the people responsible of the murder were convicted.
- *Taken captive:* A binary variable that describes if the journalist was kidnapped in an immediate period prior to the death.
- *Threatened:* A binary variable that refers to any threat to the journalist any time before death.
- *Tortured:* A binary variable that describes if the journalist was demonstrably physically tortured before of being killed. Two labels: yes, and no.

Objectives and Hypotheses

The main objective of this project is to visualize the dataset. Since we have geographical data, the principal visualization will be a map that displays the counts of deaths with respect to various variables in the data (such as by year or coverage), showing the evolution through the years.

The number of deaths is a quantitative variable through time (from 1992 to 2016), thus it could be seem as a time series. One of the objectives is to perform time series analysis to see how number of deaths vary through time, and possibly find patterns that cause this trend.

We also aim to generate a network graph, that connects Journalists to the regions they died in and the topics they were reporting. We aim to compare the different visualizations and see if one provides better representation than the others. We aim to visualize and find out regions and topics that may be more perilous.

We plan to perform some detailed analysis using PCA and Factor Analysis, to see if we can try to explain how the countries (or years) of deaths group together.

Acknowledgements

We would like to thank professor Yong-Yeol Ahn for providing us the opportunity to work on such an interesting visualization project.

We would also like to thank the Committee to Protect Journalists for collecting the data and making the data publicly available and facilitating research in the topic.

We would like to thank Kaggle for their data platform that makes accessing and downloading datasets very easy for data scientists.

References

[Adhokshaja 2015] Adhokshaja, P. 2015. Eda and animation. Kaggle Kernels.

[Balguy-Gallois] Balguy-Gallois, A. The protection of journalists and news media personnel in armed conflict. *International Review of the Red Cross* 86(853):37–67.

[Biggs 2018] Biggs, J. 2018. Python Package Index wesite.

[Committee to Protect Journalists 2016] Committee to Protect Journalists. 2016. Journalists killed worldwide since 1992. Kaggle.

[Committee to Protect Journalists 2018a] Committee to Protect Journalists. 2018a. 1321 journalists killed. Committee to Protect Journalists Website.

[Committee to Protect Journalists 2018b] Committee to Protect Journalists. 2018b. 42 journalists killed. Committee to Protect Journalists Website.

[Committee to Protect Journalists 2018c] Committee to Protect Journalists. 2018c. Committee to protect journalists. CPJ website.

[Council of Europe 2018] Council of Europe. 2018. About the platform for the protection of journalism and safety of journalists. COE website.

[geopandas 2018] geopandas. 2018. Geopandas 0.4.0. geopandas docs.

[International Federation of Journalists 2018a] International Federation of Journalists. 2018a. About us. International Fererederation of Journalists Website.

[International Federation of Journalists 2018b] International Federation of Journalists. 2018b. Journalists & media staff killed list in 2018. International Federation of Journalists Website.

[Kishore 2017] Kishore, R. 2017. Making india safe for journalists. Live Mint Website.

[Langley 2014] Langley, A. 2014. Maps reveal violence against journalists. Columbia Journal Review.

[Lara 2016] Lara, M. 2016. Exploratory analysis of cpj data tahuichimiguel. Kaggle Kernels.

[Marlowe 2018] Marlowe, L. 2018. Three years after charlie hebdo, magazines staff live in fear. the Irish Times Website.

[National Coalition Against Censorship 2018] National Coalition Against Censorship. 2018. Visualizing censorship. National Coalition Against Censorship website.

[Networkx Contributors 2018] Networkx Contributors. 2018. Software for complex networks. Networkx website.

[Plotly contributors 2018] Plotly contributors. 2018. Network graphs in python. Plotly docs wensite.

[Reporters Without Borders 2018a] Reporters Without Borders. 2018a. 2018 world press freedom index. Reporters Without Borders website.

[Reporters Without Borders 2018b] Reporters Without Borders. 2018b. Clear increase in violence against journalists in germany in 2018. Reporters Without Borders.

[Reporters Without Borders 2018c] Reporters Without Borders. 2018c. Missing journalists. Reporters Without Borders Website.

[Reporters Without Borders 2018d] Reporters Without Borders. 2018d. Protecting journalists. Reporters Without Borders Website.

[scikit-learn contributors 2018] scikit-learn contributors. 2018. sklearn.decomposition.pca. scikit-learn documentation website.

[Seelke 2018] Seelke, C. R. 2018. Violence against journalists and media workers in mexico and u. s. policy.

[Tenkanen 2017] Tenkanen, H. 2017. automating gis processes. GeoPython Auto-GIS website.

[The Hindu 2018] The Hindu. 2018. Reporters without borders urges union govt. to safety of journalists. The Hindu Website.

[UCLA Image Institute for Digital Research and Education 2018] UCLA Image Institute for Digital Research and Education. 2018. A practical introduction to factor analysis: Exploratory factor analysis. UCLA IDRE website.

[Whitaker 2017] Whitaker, N. 2017. Our efforts to help protect journalists online. Google Blog.

[wiki.gis contributors 2018a] wiki.gis contributors. 2018a. Geographic information system. wiki.GIS.com website.

[wiki.gis contributors 2018b] wiki.gis contributors. 2018b. Geographic information system. wiki.GIS.com website.

[Wikipedia contributors 2018a] Wikipedia contributors. 2018a. Charlie hebdo shooting — Wikipedia, the free encyclopedia. [Online; accessed 16-September-2018].

[Wikipedia contributors 2018b] Wikipedia contributors. 2018b. Committee to protect journalists — Wikipedia, the free encyclopedia. [Online; accessed 16-September-2018].

[Wikipedia contributors 2018c] Wikipedia contributors. 2018c. Council of europe — Wikipedia, the free encyclopedia. [Online; accessed 16-September-2018].

[Wikipedia contributors 2018d] Wikipedia contributors. 2018d. Principal component analysis — Wikipedia, the free encyclopedia. [Online; accessed 17-September-2018].

[Wikipedia contributors 2018e] Wikipedia contributors. 2018e. Reporters without borders — Wikipedia, the free encyclopedia. [Online; accessed 16-September-2018].

[Wikipedia contributors 2018f] Wikipedia contributors. 2018f. Time series — Wikipedia, the free encyclopedia. [Online; accessed 17-September-2018].

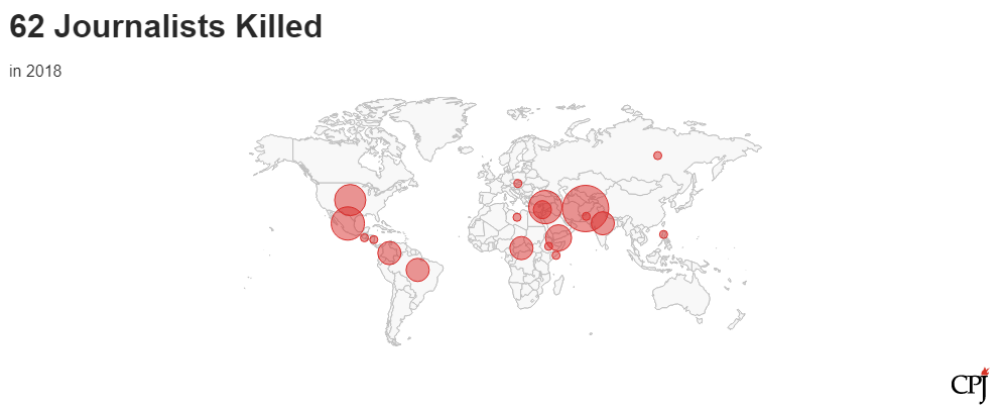

Figure 1: Map of journalists killed in 2018 (Committee to Protect Journalists 2018b)

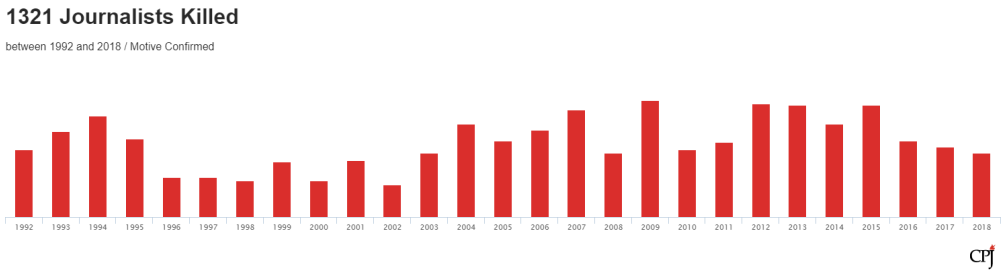

Figure 2: Histogram of journalists killed in 2018 (Committee to Protect Journalists 2018a)

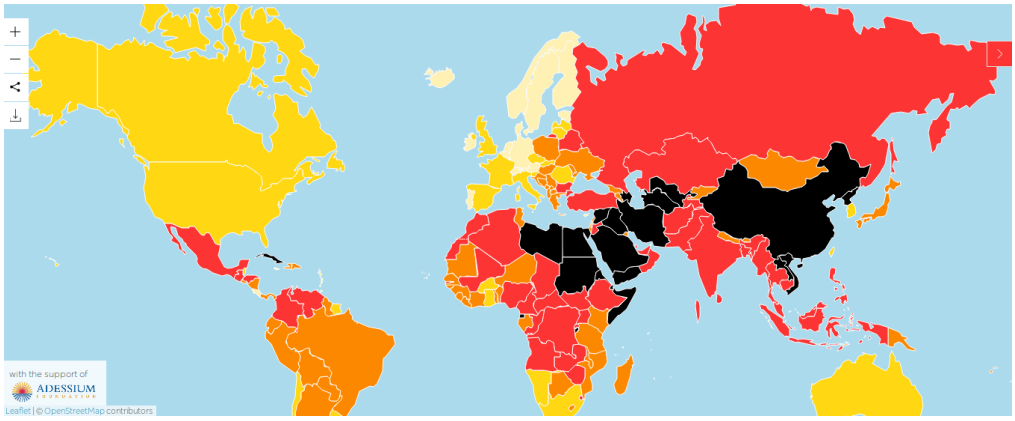

Figure 3: Freedom index across the world by RSF (Reporters Without Borders 2018a)

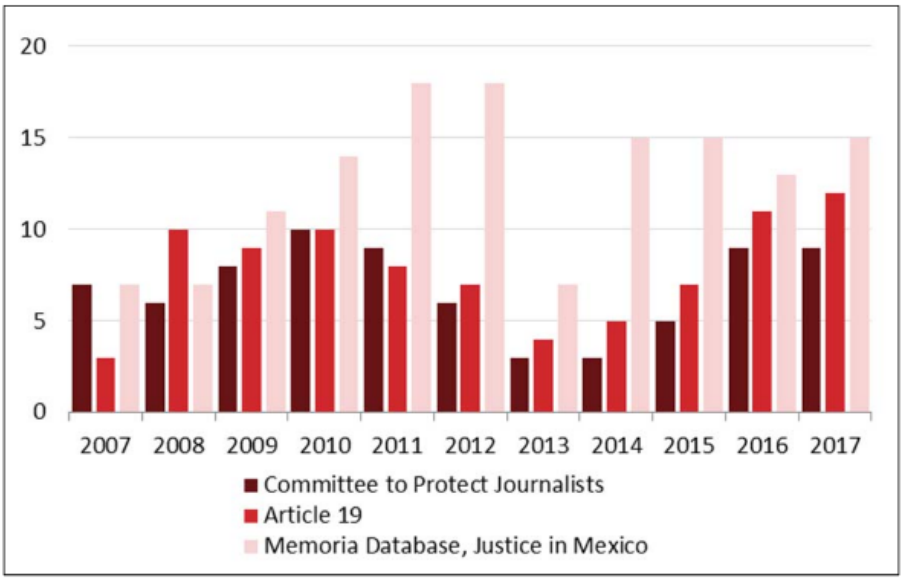

Figure 4: Homicides of Journalists ans Media Workers in Mexico: 2007-2017 (Seelke 2018)

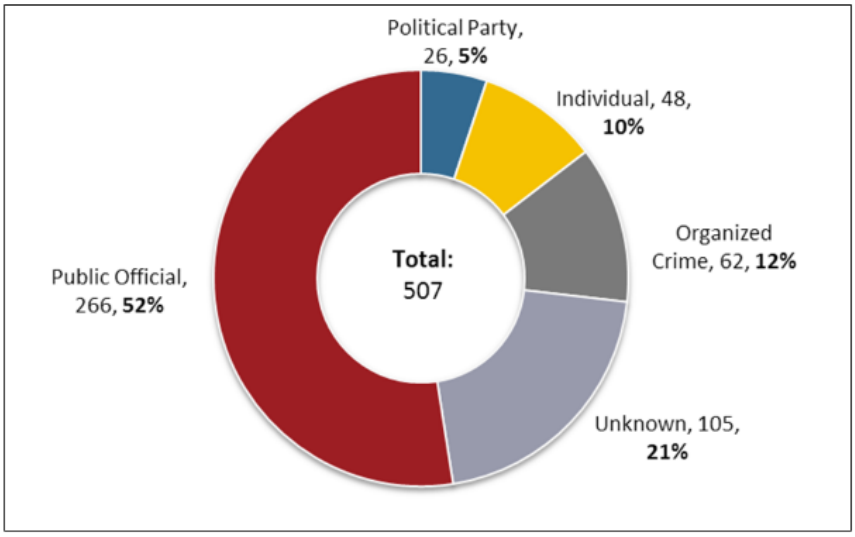

Figure 5: Perpetrators of Violence against Journalists in Mexico: 2017 (Seelke 2018)

Figure 6: Violations of Free Expressions in Europe: 2014 (Langley 2014)

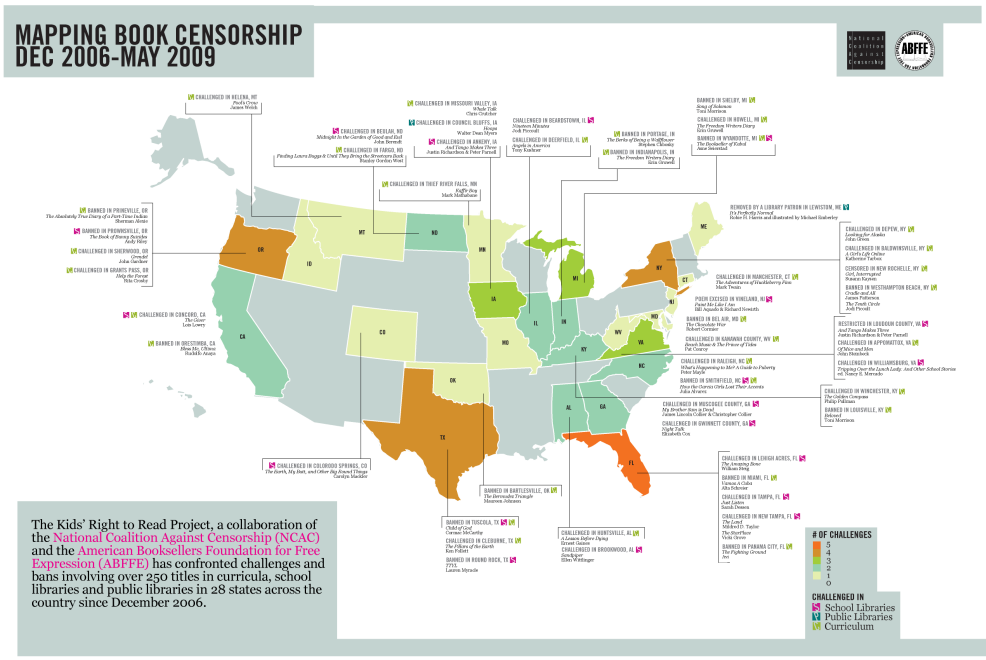

Figure 7: Book Censorship in United States (National Coalition Against Censorship 2018)

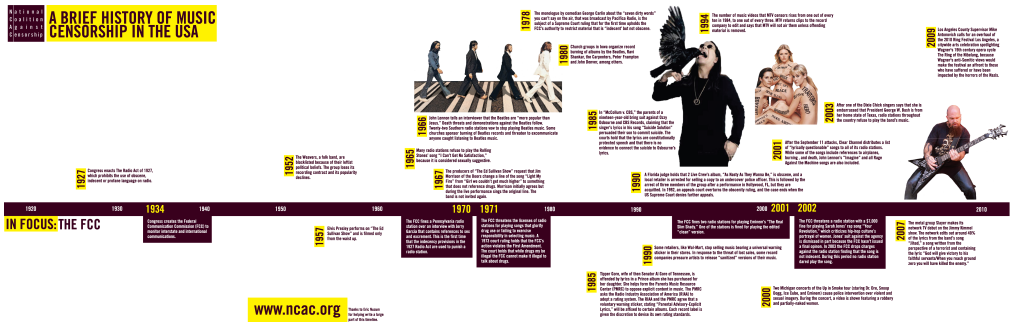

Figure 8: Music Censorship in United States (National Coalition Against Censorship 2018)