2. RELATED WORK

Recent years have seen the development of research projects aiming to tackle the complexities of researching the online media ecosystem. Probably the most comprehensive is the MediaCloud platform [4]. This open source, open data, platform collects and analyzes the news stream of tens of thousands of online sources. The system provides real time search of an archive of over 2.5 billion sentences from over 50 thousand online media sources. The motivation to build MediaCloud derived from the need for tools that offered a way for researchers and academics from the media fields, among others, to ask complex questions about the online media ecosystem.

Researchers first define the set of media sources they wish to collect. The MediaCloud platform then discovers the RSS feeds for each of the selected sources, and crawls them several times each day to collect new stories. The system downloads the HTML of each story and then extracts the primary text of the page, filtering out the accessory content. The substantive text is broken down into a set of word counts, which are subsequently analyzed and processed by a set of text analysis tools [5].

Since the inception of the MediaCloud platform, several tools have been built on top of it. For example, the MediaMeter Dashboard [6] provides a quick overview of an issue's media coverage and related trends, and the MediaMeter Focus [7] displays the global geographic coverage of distinct types of media. The system has also been used to research media controversies [8, 9], or to compare the roles of blogs and mainstream media [10].

Another relevant online news analysis system is the BBC News Labs Juicer [11]. The platform provides access to the content of several news sources, which is mainly comprised of news articles, but also includes images, video and tweets from select sources. Articles are semantically annotated with concepts based on DBPedia entries, such as people, organizations and places. The Juicer API provides endpoints for finding or retrieving concepts, get concept occurrences and co-occurrences, and full-text search for articles. Relevant tools developed on top of the Juicer API include the WAT application [12], which allows querying the content of hundreds of news sources and displays the results visually, or the NewsMap [13], a 3D globe which automatically navigates to certain regions of the world and displays stories linked to them.

We also found an interesting tool named NewsStand [14], developed for the visualization of the geographic content in news articles. The motivation behind the development of this prototype was the emphasis on "Where" a news event is taking place. To answer this query, NewsMap uses a database that associates news articles with the geographic references mentioned in them, and groups these articles into story clusters based on their textual and geographic content, placing the markers of the story clusters in an interactive map interface.

We decided to build our own set of tools because the above mentioned systems do not collect data from several online sources that we wanted to investigate, have limited access or do not provide some required features. These features include access to the sources' publication patterns data according to distinct temporal granularities, collecting the number of shares per article on Twitter and Facebook, visualization of online media sources' geographic coverage at the Portuguese district level, and the possibility of finding co-occurrences of search terms and geographic places.

3. MEDIAVIZ PLATFORM

The MediaViz platform can be described as two communicating components. The first is the storage and access system, a Ruby on Rails web application that collects, persists and manages the news data, allowing programatic access to it through a JavaScript Object Notation (JSON) API. The application includes an administrative User Interface (UI) where users can add, delete and edit the sources and feeds to be collected. All the sources' feeds are regularly parsed by a crawler module responsible for saving new articles to a PostgreSQL database. The crawler module also communicates with Twitter's and Facebook's APIs to collect the number of shares for each article. Finally, the API module provides programatic access to the data in JSON text format through several endpoints.

The second component is a client application composed by a set of distinct and self-contained visualization tools which allow the exploration and interaction with the collected data. The client application is a Single Page Application (SPA), developed using the JavaScript (JS) framework AngularJS. It communicates with the API, uses the retrieved data to generate the visualizations, and renders the UI. Regarding the visualization technology stack, two JS libraries were used: C3.js and Data-Driven Documents (D3) [15]. C3.js is a library built on top of D3 that abstracts much of the code needed to generate common graphics, such as area and line charts. For the visualizations with more strict requirements, such as maps, 'pure' D3 was employed. In this paper we will focus on three distinct visualization tools, described below.

4. MEDIAVIZ VISUALIZATIONS

4.1 Chronicle

The Chronicle tool allows users to insert multiple search terms and visualize the amount of coverage they received throughout time. The tool searches the title and summary of articles published by a large predefined group of sources (Portuguese media, International media, and Portuguese blogs). The tool's name, as well as some functionality, was inspired by the homonym tool from the New York Times Research and Development group [16]. By comparing the attention granted by a large set of media sources to distinct terms, the Chronicle tool displays which topics have been and currently are under the media's spotlight, and how those dynamics change throughout time and between distinct sets of sources. While the reasons why certain topics receive a substantial amount of coverage might be clear due to its evident news value [17], there are cases where it might not be immediately obvious. Thus, to further understand why a certain topic received a determinate amount of coverage, users can click the chart and see the linked articles.

We will exemplify by comparing the coverage of two terms between a set of traditional online Portuguese news sources and a set of Portuguese blogs. The two terms compared were: 'Charlie Hebdo', the french satirical magazine which was the target of a major terrorist attack on 7 January 2015, and 'Ronaldo', which is the name of the Portuguese football player who won the FIFA Ballon d'Or Award in January 12

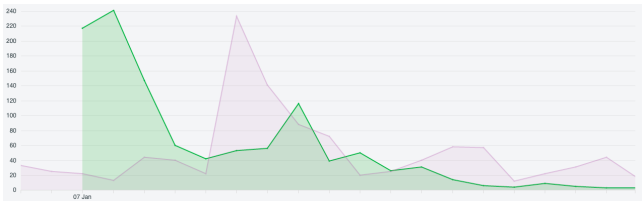

Figure 1: Chronicle - Volume of articles published by all Portuguese sources by day mentioning 'Charlie Hebdo' (dark) and 'Ronaldo' (light).

Figure 2: Chronicle - Volume of articles published by all Portuguese blogs by day mentioning 'Charlie Hebdo' (dark) and 'Ronaldo' (light).

2015 (attributed every year to the best performing football player from the previous calendar year).

Figure 1 shows the volume of articles for each of the selected terms published by all Portuguese traditional online news media. It is possible to see how the number of articles published daily follows the events: the term 'Charlie Hebdo' (dark) was under the media's focus on 7 and 8 January 2015, and the term 'Ronaldo' (light) received a similar amount of attention on 12 January 2015. When examining the Portuguese blogs, a different behavior is discernible. Figure 2 displays the number of articles containing the same search terms published by all Portuguese blogs in the system. It shows that 'Charlie Hebdo' (dark) received a fair amount of attention on 7, 8 and 9 January 2015, while 'Ronaldo' (light) was barely mentioned on 12 January 2015. It should be mentioned that, while the group of Portuguese news media includes every major online outlet, the set of blogs is more limited and comprised mainly of sites covering topics such as politics and economy.

Even though, this example highlights how the Chronicle tool can be useful to examine coverage dynamics or to identify how different types of sources attribute importance to distinct issues.

4.2 Flow

The Flow tool allows users to visualize and compare the article publication patterns of online sources. Users can view the data according to distinct temporal granularities: in 24-hour, weekly, and monthly cycles. It also provides quick access to predefined intervals (everything, last day, last seven days, last 30 days) and customizable date intervals. In addition to the volume of articles published by source, it is also possible to see the aggregated number of shares on Twitter, Facebook, or both. The Flow tool is mainly exploratory: through the visual comparison of publishing and sharing behaviors, it can help unearth interesting research questions. As mentioned by Aigner et al. [2], the exploration and analysis of time-oriented data requires appropriate visual and analytical methods. Our approach consists of representing time as a continuous flow, instead of a series of discrete events.

In order to have comparable results, all dates are parsed according to the Coordinated Universal Time (UTC) time standard. In addition to the absolute count of articles, users can select the 'Relative' mode, which displays the percentage of articles published relative to the source's total articles. This permits more discernible comparisons, particularly when there's a disparity between the sources' total article count. The comparison between sources helps highlighting differences in publication and sharing patterns (even when the sources are within the same timezone): while some display a regular flow (e.g., continue publishing during the dawn), others leverage certain schedules (e.g., publishing spikes close to lunch and diner time).

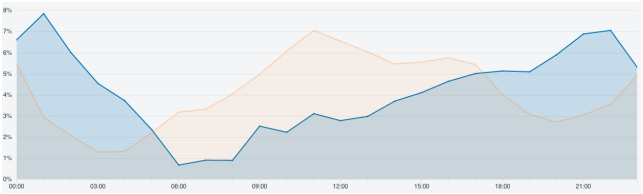

Figure 3: Flow - Relative volume of articles published in a 24-hour cycle by the New York Times (dark) and the BBC (light).

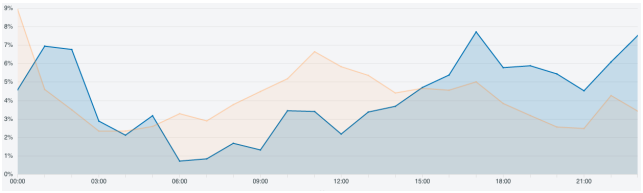

Figure 4: Flow - Relative volume of aggregated share counts on Twitter and Facebook for articles published in a 24-hour cycle by the New York Times (dark) and the BBC (light).

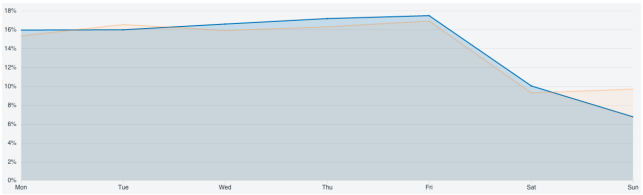

Figure 5: Flow - Relative volume of articles published in a weekly cycle by the New York Times (dark) and the BBC (light).

Let's exemplify with a user that wishes to compare the publishing patterns of the the New York Times and the BBC in a 24-hour cycle. After selecting these two sources in the source selection dropdown, the user then selects the 'Hour' and the 'Relative' options in the UI. This results in the visualization shown in Figure 3, which displays how the New York Times (dark) and the BBC (light), even taking into

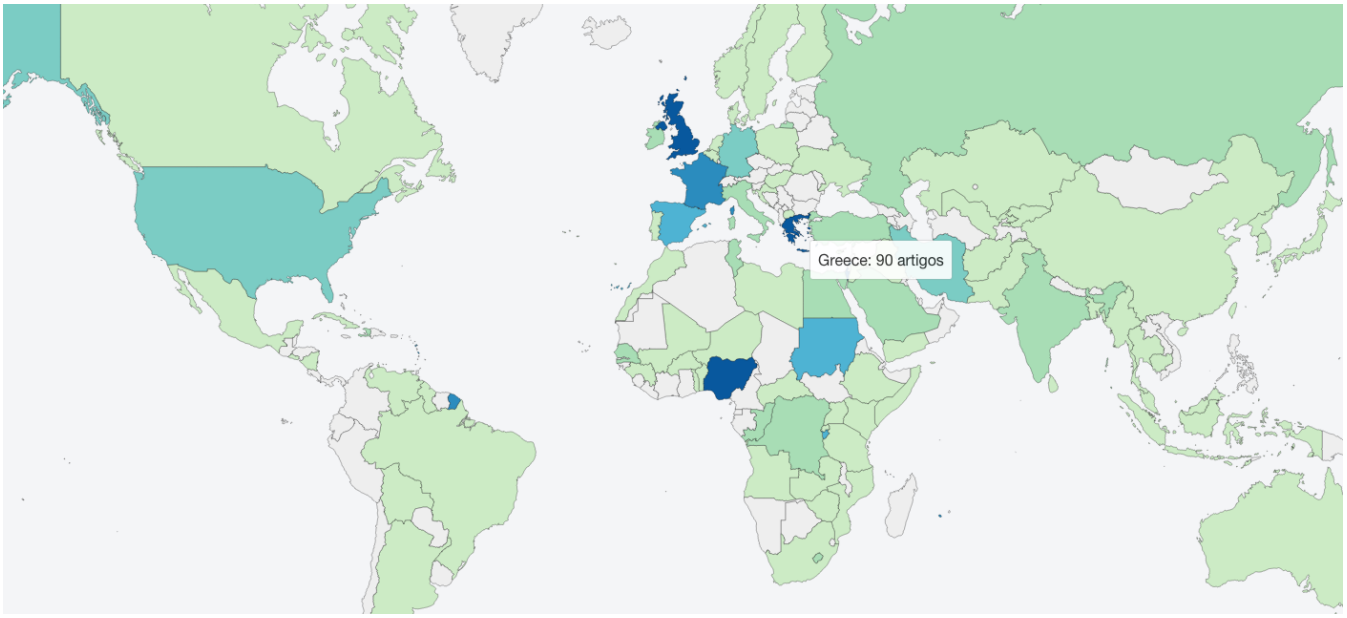

Figure 6: NewsMap - Co-occurrences of the term 'elections' and countries' names in articles published by all international sources between 1 January and 1 June 2015.

account time zone differences, have very distinct publication profiles when comparing the relative volume of articles published in a 24-hour cycle. Let's say that the user now aims to see if the share behavior on Twitter and Facebook somehow reflects, or not, the news publishing patterns of these two sources. The user selects the 'Twitter + Facebook' option, and the visualization is rendered (Figure 4). Figure 4 shows that the share pattern for the BBC (light) closely mirrors the publication pattern displayed on Figure 3 (i.e., there's a positive relationship between the volume of articles published and the number of shares). The biggest difference happens at 00:00 UTC, when the volume of shares is considerably higher than the number of articles published, in relative terms. On the other hand, the New York Times (dark) displays a more irregular pattern. However, when comparing both sources' publications in a weekly cycle, they exhibit very similar patterns, as shown in Figure 5. In order to better understand the reasons behind certain patterns, the user can click the data points to see the associated articles. Each article is comprised by its title, summary, publication date, a link to the source URL, and the share counts on Twitter and Facebook.

4.3 NewsMap

The NewsMap tool displays a map of the world, subdivided by countries. Every time a source is selected, the map redraws and colors each country according to the number of articles that mention it. In addition to the world map, users can visualize a map of Portugal, subdivided by administrative units, that follows the same design and interaction principles. This type of visualization, known as a choropleth map, is an efficient way to display the variation of a variable of interest across a geographic region [18]. The choice of the color scheme can have a significant impact in how viewers perceive the information displayed [18]. Thus, the ColorBrewer tool [19], which provides a set of research-backed color schemes for thematic maps, was employed to maximize clarity and enhance comprehension.

In terms of insights, this tool allows users to visualize a source's (or group of sources) geographic coverage, both at the international and national level. Users can also input text to find co-occurrences of the introduced search term and the countries' or Portuguese subdivisions' names. Moreover, users can select a date range to limit the search to articles published between the specified period. Let's assume a user wants to see the countries most mentioned in conjunction with the term 'elections' between between 1 January 2015 and 1 June 2015, on international (i.e., non-Portuguese) media. After selecting the option 'All international' in the sources selection dropdown, entering the term 'election' in the search box, and setting the interval for the desired period, the NewsMap tool displays all countries co-mentioned with the term 'elections' colored according to the volume of articles (darker meaning more articles), as shown in Figure 6.

The map shows that Greece (which held a legislative election on January 25 2015) was the most covered country, with 90 articles, followed by Nigeria (general election on 28-29 March 2015; 85 articles), Israel (legislative election on 17 March 2015; 67 articles), and the United Kingdom (general election on 7 May 2015; 57 articles). The visualization also displays that France (departmental elections on 22 and 29 March 2015; 34 articles), Spain (municipal elections on 24 May 2015; 26 articles) and Sudan (general elections on 13-16 April 2015; 24 articles) received a significant amount of coverage. Finally, Burundi also appears linked with a high international media's focus (24 articles) due to a political crisis related with the president's decision to run for a third term in office in the presidential election scheduled to 26 June 2015.

In short, by using the NewsMap tool as exemplified above, it is possible to quickly identify which country's elections

were most scrutinized by international news media according to the number of articles, which can be seen as an indicator of relative importance. Additionally, users can click the countries and see all the linked articles to better understand the underlying reasons for the received coverage.

5. CONCLUSION

The work presented in this paper represents the first iteration of the MediaViz platform. The next version is currently under development and involves refinement of the UI based on user research, exploration of new visualization tools and new analytical approaches. The integration of robust search and content analysis tools in the infrastructure is also planned. This will allow for more complex content analysis tasks, such as named entity extraction. It is also expected the implementation of data export capabilities, both in text (e.g., CSV, JSON) and image formats (e.g., PNG, SVG, PDF). We described how the visualization tools presented aim to assist the study of the online media ecosystem by putting news data under different lens to permit distinct analysis and insights. These include the flow of publications throughout time, and how the media's spotlight changes its focus based on the news source, the theme and the geographic region. We believe they can help researchers, academics, and professionals from the media field to gain a deeper understanding regarding the behavior of news sources in the online space.

6. ACKNOWLEDGMENTS

This work was partially supported by SAPO Labs / U.Porto, financed by PT Comunicações and SIBILA Project - NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

7. REFERENCES

[1] Da Keim. Visual exploration of large data sets. *Communications of the ACM*, 44(8):38–44, 2001.

[2] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[3] Ji Soo Yi, Youn-ah Kang, John T Stasko, and Julie a Jacko. Understanding and characterizing insights: How Do People Gain Insights Using Information Visualization? *Proceedings of the 2008 conference on BEyond time and errors novel evaLuation methods for Information Visualization - BELIV '08*, page 1, 2008.

[4] About Media Cloud | Media Cloud. http://mediacloud.org/about/.

[5] Jason Chuang, Sands Fish, David Larochelle, William P. Li, and Rebecca Weiss. Large-Scale Topical Analysis of Multiple Online News Sources with Media Cloud. *NewsKDD*, August 2014.

[6] Dashboard - MediaMeter. https://dashboard.mediameter.org/#demo.

[7] Media Meter: Focus. http://focus.mediameter.org/.

[8] Erhardt Graeff, Matt Stempeck, and Ethan Zuckerman. The battle for 'Trayvon Martin': Mapping a media controversy online and off-line. *First Monday*, 19(2), January 2014.

[9] Yochai Benkler, Hal Roberts, Robert Faris, Alicia Solow-Niederman, and Bruce Etling. Social Mobilization and the Networked Public Sphere: Mapping the SOPA-PIPA Debate. *SSRN Electronic Journal*, July 2013.

[10] Bruce Etling, Hal Roberts, and Robert Faris. Blogs as an alternative public sphere: The role of blogs, mainstream media, and tv in russia's media ecology. *Berkman Center Research Publication*, (2014-8), 2014.

[11] BBC - The Juicer: Semantic Entity Extraction - BBC News Labs - Connected Studio. http://www.bbc.co.uk/partnersandsuppliers/connectedstudio/newslabs/projects/juicer.html.

[12] BBC News Labs - Who's talking about what. http://wat.bbcnewslabs.co.uk/.

[13] BBC News Labs World News Map. http://newsmap.bbcnewslabs.co.uk/.

[14] Benjamin E. Teitler, Michael D. Lieberman, Daniele Panozzo, Jagan Sankaranarayanan, Hanan Samet, and Jon Sperling. Newsstand: A new view on news. In *Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '08, pages 18:1–18:10, New York, NY, USA, 2008. ACM.

[15] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2301–2309, 2011.

[16] NYT Chronicle. http://chronicle.nytlabs.com/.

[17] Johan Galtung and Mari Holmboe Ruge. The structure of foreign news the presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, 2(1):64–90, 1965.

[18] Cynthia A Brewer, Alan M MacEachren, Linda W Pickle, and Douglas Herrmann. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, 1997.

[19] Mark Harrower and Cynthia A Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.