

---

This is an electronic reprint of the original article.  
This reprint may differ from the original in pagination and typographic detail.

Koivunen-Niemi, Laura; Masoodian, Masood  
**Visualizing narrative patterns in online news media**

*Published in:*  
Multimedia Tools and Applications

*DOI:*  
[10.1007/s11042-019-08186-9](https://doi.org/10.1007/s11042-019-08186-9)

Published: 11/10/2019


*Document Version*  
Publisher's PDF, also known as Version of record

*Published under the following license:*  
CC BY

*Please cite the original version:*  
Koivunen-Niemi, L., & Masoodian, M. (2019). Visualizing narrative patterns in online news media. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-019-08186-9>



# Visualizing narrative patterns in online news media

Laura Koivunen-Niemi<sup>1</sup> · Masood Masoodian<sup>1</sup> 

Received: 23 November 2018 / Revised: 25 July 2019 / Accepted: 6 September 2019 /

© The Author(s) 2019

## Abstract

News media play an important role in shaping social reality, and their multimedia narrative content, in particular, can have widespread repercussions in the public's perception of past and present phenomena. Being able to visually track changes in media coverage over time could offer the potential for aiding social change, as well as furthering accountability in journalism. In this paper, we explore how visualizations could be used to examine differences in online media narrative patterns over time and across publications. While there are existing means of visualizing such narrative patterns over time, few address the aspect of co-occurrence of variables in media content. Comparing co-occurrences of variables chronologically can be more useful in identifying patterns and possible biases in media coverage than simply counting the individual occurrences of those variables independently. Here, we present a visualization, called *time-sets*, which has been designed to support temporal comparisons of such co-occurrences. We also describe an interactive prototype tool we have developed based on *time-sets* for analysis of multimedia news datasets, using an illustrative case study of news articles published on three online sources over several years. We then report on a user study we have conducted to evaluate the *time-sets* visualization, and discuss its findings.

**Keywords** Temporal visualization · Co-occurrence visualization · Multimedia content · Time-sets · Visual design · Journalism

## 1 Introduction

Many of the, now iconic, visualizations designed by the likes of William Playfair, Charles Joseph Minard, and Florence Nightingale (for examples of their work see [1]) were originally created and used for raising public awareness of social, political, economic, and other such issues. Similarly, since the early days of widespread newspaper publications in the late 1800s, visualizations – particularly statistical graphs [11] – have been used along with text

---

✉ Masood Masoodian  
[masood.masoodian@aalto.fi](mailto:masood.masoodian@aalto.fi)

Laura Koivunen-Niemi  
[laura.koivunen-niemi@alumni.aalto.fi](mailto:laura.koivunen-niemi@alumni.aalto.fi)

<sup>1</sup> School of Arts, Design and Architecture, Aalto University, Espoo, Finland

and imagery in reporting news. These days, the use of various forms of visualizations and “infographics” is common in most print or online newspapers. The use of visualizations in these cases is mainly to make some type of data understandable to the reader, which, with the ever-increasing complexity and abundance of available data, is becoming evermore crucial in making such data more accessible.

Therefore, although the use of visualizations is no longer novel in journalism, its full potential is, however, far from being realized. Particularly its applications in data journalism – investigative data-laden research meant for public consumption – could be further utilized to help bring about political and social change. To that end, the potential persuasive powers of visualizations as a key element of data journalism could shed light on social phenomena, many of which have proved to be divisive within society in recent times. For instance, viewpoints even within the European Union often diverge greatly on any given subject, particularly on immigration into the region over the past few years – it has for example become painfully apparent that there are some distinct biases and preconceptions about immigrants as people, their motives in immigrating, their impact on society, and so on [19]. The question that arises is, how could people from a fairly constrained geographical area (in this case EU), with relatively similar cultural backgrounds, hold such opposing views on an issue? Could the differences in public opinion be, to a large extent, the result of how media in different countries have covered such social or political issues? More importantly, how could a group of people with such dissimilar views ever discuss these issues in a constructive manner, without understanding the reasons behind such different viewpoints?

In this paper, we propose the use of temporal visualizations for examining differences in media narrative patterns over time and across publications. The aim in this is not to expose particular biases in media, nor to imply direct causality between media events and changes in social narratives, but simply to report findings in an effective way that could be used as an additional tool for media analysis and discussion on the subject.

Visualizations in this sense can contribute to discourse about media use, and furthermore to the field of media research itself. Providing media researchers with the ability to quickly identify patterns in a visualization of the collected data could be immensely valuable. For example, pointing out when certain topics began appearing together in common discourse could signify a major shift in narratives and public opinion in general – the ease of tracking variable flow through data could be considerably faster with an appropriate visualization tool. Traditional methods of communicating media analysis information fall short in their ability to simplify and communicate patterns such as this in large datasets [17]. The key specifically lies in the ability to automate the processing of large and complex datasets into a visual form which is more comprehensible. This task could, of course, be approached from several different angles.

Here, we describe the *time-sets* visualization [28], which uses a modified form of *linear diagrams* [13], to allow comparing co-occurrence of multimedia content variables chronologically across selected publications. As a case study, we then illustrate the use of this technique to analyze the relationships between articles on certain topics published by several online news sources over a period of time, and present an interactive prototype visualization tool we have developed to support such analyses of multimedia online news datasets. Finally, we report on a user study we have conducted to evaluate the *time-sets* visualization. We conclude with a discussion of the findings of this study and some of its implications for future improvements.

It should also be noted that, while the work reported here is related to our earlier work [28], in this paper we focus more on the application of our *time-sets* visualization in the area of journalism, rather than looking at its potential as a new set visualization technique

– which was our aim in [28]. As such, here we also review other existing media content visualization methods, more fully describe the *time-sets* visualization prototype we have developed along with a range of interactive tools, provide more detailed journalism-related case study examples using a larger set of *time-sets* visualizations, and present an evaluation of the *time-sets* visualization which attempts to gauge its effectiveness in supporting such journalistic tasks.

## 2 Review of media visualizations

While the study of large datasets of text using natural language processing methods is becoming more widespread in many disciplines, methods of visualizing these datasets have not yet been explored to their full potential. In the field of journalism, in particular, these types of visualizations hold great promise as a tool for media oversight and analysis. Being able to visually examine changes in media coverage over time could be an invaluable tool for tracking shifts in media narrative tone, not to mention the public's perception of various cultural phenomena. In this section, we present a survey of the most common applications of temporal visualizations in contemporary journalistic practice and comparative media studies.

Formal media studies have traditionally opted to present the data, and findings amassed during the study, in the most straightforward form possible. The content herein can in itself be very varied and can include numerical word frequency [26], topic occurrence [16], news framing [33], imagery choices [15], degree of perceived news credibility [18], geographic origin [8], source attribution [10], or any number of other variable occurrences during a certain time period. However, despite the large range of possible variables, the findings from these types of studies are almost invariably presented in tables [17], or using simple graphs such as bar charts [22] or line charts [12].

This unvaried use of relatively simple visual mechanisms is to some extent understandable, since the objective is arguably to only present the gathered data, rather than a desire to analyze the dataset through novel forms of visualization. The visualization in such cases is not meant as a means to an end – i.e., not a tool for research, but simply a receptacle for housing the data, or presenting already completed analysis. While tables and simple graphs may suit this purpose reasonably well, the main limitation of these more conventional visual methods is that they offer a relatively shallow view of the underlying dataset, and given enough variables, they can become increasingly unwieldy for the viewer to understand. Easily identifying patterns over time, especially patterns in variable use correlation over time, proves to be considerably more cumbersome as well in these types of charts. More complex scatter plots [6] and box plots do feature in a subset of media studies which attempt to address sentiment and correlation in news content, but these seem to be in the minority and lean further into computational linguistic studies than media studies in themselves. Similarly, other more advanced visualizations for text analysis also mainly target linguistic [25, 32] or conversation analysis [3, 4] use cases.

In comparison, the types of visualizations found in popular culture and mass media more often attempt to scrutinize phenomena through visualizations, and make use of a wider variety of visual methods. Furthermore, in contrast, these visualizations – like the popular culture content streams they persist in – tend to have a perspective or agenda [27], and aim to convince the reader through the presentation of evidence in various forms. The reasons for this dichotomy with formal media studies could perhaps stem from a larger news agency staff, funding for graphic artists, or even less rigid traditions in publication standards –

though most likely from a need to communicate, primarily visually, to a less specialized general audience in novel ways.

Some notable visualizations commonly used in popular online media news comparisons include linear bubble charts<sup>1</sup>, grouped column charts<sup>2</sup>, dot plots<sup>3</sup>, as well as simpler line charts for tracking word-use over time<sup>4</sup>. There are, however, a number of issues with the way, for instance, line charts are used in the popular media, particularly for showing word or topic occurrences. For example, charting incidences of a variable over time on a line chart implies continuity and causality in the data from one period to the next. Similarly, charting several variables in this manner on the same axis may imply that the variables have some (perhaps causal) relationship. For example, it would be easy for a viewer to draw a faulty parallel by thinking that if, for instance, words *A* and *B* both increase at the same time, then there must be a significant connection between them, and possibly a common catalyst for their changes [34]. While displaying their line graphs separately would avoid the problem of overtly implying correlation between the different line trajectories, as well as providing a common baseline for comparison, many users often fail to normalize the axis for comparison, causing other possible misconceptions about different variables' comparative scale and significance.

There are of course a plethora of other interesting variations to the basic word count and occurrence visualizations, including ternary plots of comparative network topic reporting<sup>5</sup>, stacked bar charts of exhaustive debate transcript coverage<sup>6</sup>, overall narrative sentiments in 100% stacked bar charts<sup>7</sup>, and many more. However, even though many of these visualizations are being utilized in these settings as exploratory tools, few, if any, of them support visualization of co-occurrence of variables, and none of them across the time dimension.

The addition of features to allow visual comparison of variable co-occurrences could be used to draw out unseen facets of a news dataset and help viewers perceive any existing temporal trends. A visualization employing such features could be used to analyze any variety of multimedia content, and have applications in a broader socio-cultural context as a means of tracking sentiment flow over time, within and between news sources, as well as promoting objectivity and accountability in journalism. We have, therefore, designed the *time-sets* visualization to support visual comparison of such variable co-occurrences in multimedia news datasets.

### 3 Time-sets visualization

Our review of existing visualizations, as described above, identified that any potential visualization needs to cope with large volumes of temporal multimedia data that collections of online news articles would contain. This analysis led us to decide to base the design of *time-sets* on set visualizations [28], which are capable of dealing with large collections of items. In the case of *time-sets*, the basic idea would be to consider a variable (e.g., a specific topic

<sup>1</sup> See Mo Dataviz articles by themes at <http://dataviz.mo.be/>

<sup>2</sup> <https://medium.com/jsk-class-of-2018/counting-words-in-sotu-speeches-f6ca32d0e87>

<sup>3</sup> <http://www.pewglobal.org/fact-sheet/news-media-and-political-attitudes-in-the-united-kingdom/>

<sup>4</sup> <https://books.google.com/ngrams>

<sup>5</sup> <https://pudding.cool/2018/01/chyrons/>

<sup>6</sup> <https://www.theatlantic.com/politics/archive/2016/09/debate-recaps-cable-news-clinton-trump-fox-msnbc-cnn/502223/>

<sup>7</sup> <http://www.journalism.org/2017/10/02/a-comparison-to-early-coverage-of-past-administrations/>

or keyword) to be a set, and then assign each news article which includes that variable as a member of that particular set. A group of variables (i.e., sets) could then be compared across time to show any interesting patterns which may exist in how those news articles deal with the selected variables, together or on their own.

In this manner, *time-sets* could be utilized, for instance, for comparing topic and word choices used by different news sources over time. This would make it possible to examine chronologically how topic choices changed and varied between news sources – in short, how media narratives changed over time in a particular subset of news sources. It could be extrapolated as well, that this form of visualization could track whether and when a coercive topic, for instance, was introduced into public discourse. Clearly, while the particular reason for the concurrence of certain topics or words would not in itself be explained by the visualization, nonetheless, the visualization could provide a tool for comparing the underlying data, which could then be further analyzed – perhaps using other tools or complimentary visualizations – to determine any potential causality.

### 3.1 Set visualizations

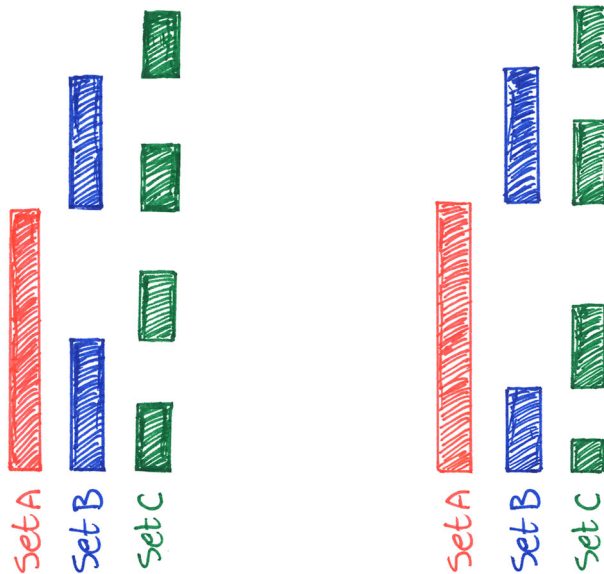
Allsallakh et al. [2] provide a comprehensive review of set visualizations, and categorize them into several groups based on their visual characteristics and the type of tasks they can support. Some of these visualizations aim to show the relationships between individual elements of the sets of interest, while others *aggregate* the set elements into their respective sets and then show the relationships between the sets themselves. These *aggregation-based* techniques are highly scalable in the number of set elements they can represent [2], and therefore, they are better suited for tasks dealing with sets of large cardinality (set size), which is very important for the tasks we envisage for *time-sets*.

*Venn* and *Euler diagrams*, and their numerous variations [29], are perhaps the most widely known *aggregation-based* set visualizations, which use the visual element of area [5] to represent sets and their relationships.

These area-based visualizations, however, do not generally show the cardinality of the individual sets or the cardinality of their intersections very effectively [13]. Even when area is used in visualizations to represent size (e.g., set cardinality), making size comparisons using area is often difficult [9]. This task becomes even more challenging when the areas to be compared are further away from each other in the visual space. As such, if representation of the time dimension, for instance, requires separation across visual space, area-based visualizations can become less than suitable for representing changes across time – a factor that we had to take into account in designing *time-sets*.

An alternative to these area-based diagrams is linear diagrams [13, 14], which have been shown to be superior for a range of visual tasks [7, 30]. Figure 1(left) shows an example of a linear diagram, representing three sets and their relationships. Linear diagrams are visually simple in their representation, and show the intersections between two or more sets through the use of parallel line segments. For example, at the bottom section of Fig. 1(left) the parallel segment of the three lines (red, blue, and green) represents the intersection between the three sets ( $A \cap B \cap C$ ), while the segment above it with two parallel lines (red and blue) represents the intersection between the two corresponding sets ( $A \cap B$ ). It should also be noted that linear diagrams are generally drawn horizontally, but here we have drawn the set lines vertically for the ease of comparisons with other illustrative examples we will provide in the rest of this paper.

Although the use of the visual element of line length [5] to represent cardinality in linear diagrams would be an obvious choice, and has previously been suggested [13, 24], it has



**Fig. 1** Examples of linear diagrams showing three sets and their relationships, without representing their cardinalities (left), and using line lengths to represent their cardinalities (right)

not been widely adopted or utilized in existing visualizations. In most cases, the drawing algorithm used for creating linear diagrams aims to create the most compact, and the least segmented, version of the diagram using multiples of a basic (shortest) unit of line length for each intersection, without representing cardinalities. Figure 1(right) shows a modified version of the linear diagram shown on the left, to represent the cardinalities of the three sets and their relationships using the individual line segment lengths.

The inherently *linear* nature of the linear diagrams, combined with the fact that the use of line lengths supports better comparisons, makes linear diagrams more suitable for use as the basis of a temporal visualization, since time itself is also most naturally represented linearly [1].

Finally, it should also be noted that there are a number of other types of set visualizations which display the cardinality of sets and their relationships using lines, in addition to their main visual representations. For instance, *UpSet* [21] and *OnSet* [31] show cardinalities using lines (bar charts) alongside the matrix of set intersections. However, such *matrix-based* visualizations are generally limited in the representation of set relationships [2].

### 3.2 Design of time-sets

The basic design of *time-sets* visualization aims to support temporal comparisons of sets and their relationships, by representing their cardinalities and relationships across time. Based on the above review, we have adopted the representation of linear diagrams for this purpose, and have utilized line lengths to show set cardinalities. We have also adopted a vertical orientation for representing the line segments, as opposed to the usual horizontal orientation, so that we could use the horizontal  $x$ -axis to represent the time dimension in its more conventional orientation.

A temporal version of linear diagrams is generated by placing the combinations of the sets of interest along a horizontal time axis, where at each point on the time axis sets are shown with their respective cardinalities and existing relationships. Figure 2 shows an example of the *time-sets* visualization. In this figure, three sets  $A$ ,  $B$ , and  $C$  are shown at three different time points  $t_1$ ,  $t_2$ , and  $t_3$ .

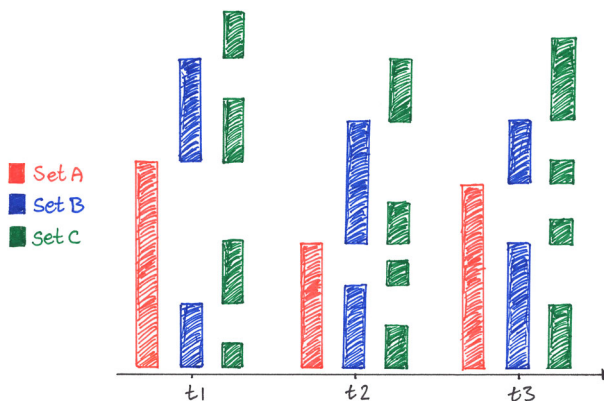
The use of the line length visual element, and the placement of the individual linear diagrams along a shared time axis, facilitate making a number of visual comparisons in support of activities related to identifying set cardinalities and relationships (e.g., intersections).

Some of the comparisons, which are based on lines aligned horizontally at the bottom, are the easiest and the most accurate to make. This is because they share a common axis, and according to the elementary perceptual tasks defined by Cleveland and McGill [9] such tasks are the easiest. For example, in Fig. 2 one can easily make an accurate judgement of the variations in cardinality of set  $A$  at different points in time (e.g., it is the smallest in  $t_2$ ). Similarly, it is possible to see the variations in the size of the intersection between the three sets ( $A \cap B \cap C$ ) at different times (e.g., it is the smallest in  $t_1$ ).

Other tasks which require visual positioning along non-aligned axes are also effectively supported, though slightly less easily [9]. For example, in Fig. 2 one can make an easy comparison of the size of the intersections between sets  $B$  and  $C$  ( $B \cap C$ ) at three different points in time (e.g., it is the smallest in  $t_3$ ).

An important point to note here is that, since the order in which the selected sets are drawn makes a big difference in terms of the number of line segments into which they are divided visually, this choice has to be made carefully to support the tasks to be performed. For instance, one might decide to draw the set with the largest cardinality, or the highest number of intersections, first to make comparisons of other sets against it easier. Rodgers et al. [30] discuss several issues related to this, and suggest several options which they then compare empirically.

We would also like to note that *time-sets* would be more suitable for the types of tasks in which the number of sets to be compared over time would be small (e.g., 3 or 4). In such cases, even if the time points along which comparisons are made are larger, detection of visual patterns based on line lengths, colors, and placement would be possible. This makes *time-sets* useful for our proposed tasks here, in which the number of online news sources



**Fig. 2** A *time-sets* diagram representing the cardinalities of three sets and their intersections at three points in time



that would be compared, as well as the number of terms (e.g., keywords) to be compared would be small, but the number of individual articles and the duration of time over which comparisons would be made could be large.

In other cases where the number of publications (i.e., sets) are much larger, alternative methods of visualization need to be investigated, or some automatic pre-processing of data needs to be carried out – for instance, using data mining techniques – to reduce the number of sets to be investigated visually by a human analyst.

## 4 Visualization prototype

We have developed a visualization prototype based on *time-sets* to support visual analysis of the relationships between online multimedia news articles, in terms of the co-occurrence of selected variables contained therein. The elements chosen for making comparisons could be related to any multimedia content. Examples include, not only words and topics choices from the textual content, but also characteristics of images, videos, or audio content, based on the elements they include (e.g., image composition, color, etc.) or their metadata (e.g., sources, authenticity, etc.).

We have tested this prototype for comparing words and topic choices in online news articles using a sample dataset we have created. In this dataset, we have included entire news articles from three leading news sources in the United Kingdom<sup>8</sup> from 2010 to 2015.

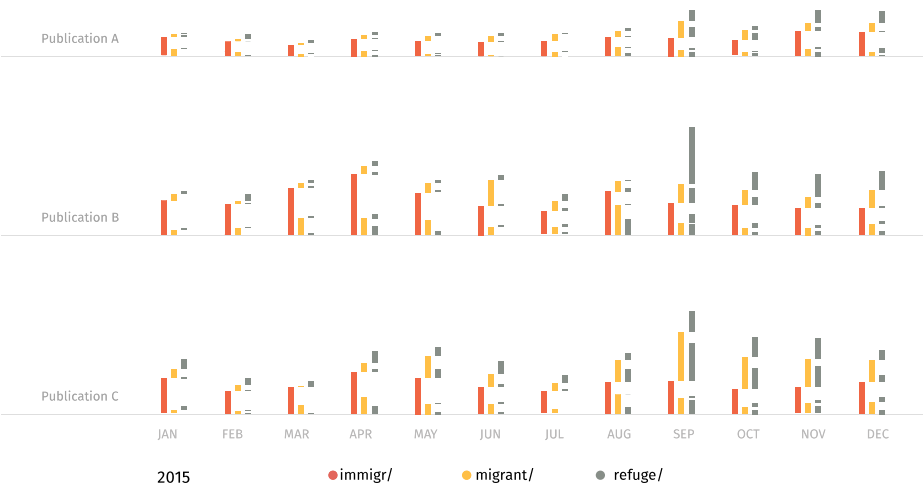
The choice of news sources and the subset of their articles to be analyzed does in itself introduce a possible element of bias into the data, and thereby, into what is being visualized. Although for our demonstrative purposes here such potential biases are not critical, it is, however, important to specify how the source material was selected. For this particular dataset, we decided to include articles categorized by each news source as “news”, “UK news”, “world news”, or “politics”, and excluded all other categories such as sports, entertainment, culture, etc. The other categories were primarily excluded in the interest of keeping the sample dataset at a manageable size for the purpose of the visualization, and keeping the articles as much on a news- and event-oriented trajectory as possible.

From this sample dataset, we extracted articles from each publication source month by month. We then queried each article using our chosen keywords (i.e., our sets), which determined the membership of that article in each of those sets. These set memberships were subsequently used to determine the specified set relationships (e.g.,  $A \cap B \cap C$ ,  $A \cap B$ , etc.) which could be visualized using *time-sets*.

The choice of the words to be queried is clearly task dependant. Here, we wanted to investigate, for instance, how different online news sources might be reporting news related to political issues of “immigration”, “migration”, and “refuge”. As these words may occur in different forms in each article, rather than searching just for these specific words, we searched for their individual stems and their inflected variants.

Figure 3 shows an example of *time-sets* using the articles published in 2015. In this visualization, each month has its corresponding linear diagram of word occurrences, where articles with any of the variations of each selected word are members of the set representing that word (e.g., articles related to “immigration” are in the set shown in red). The sets

<sup>8</sup>To void affecting public opinion about these publications unnecessarily, we will not refer to their names here.



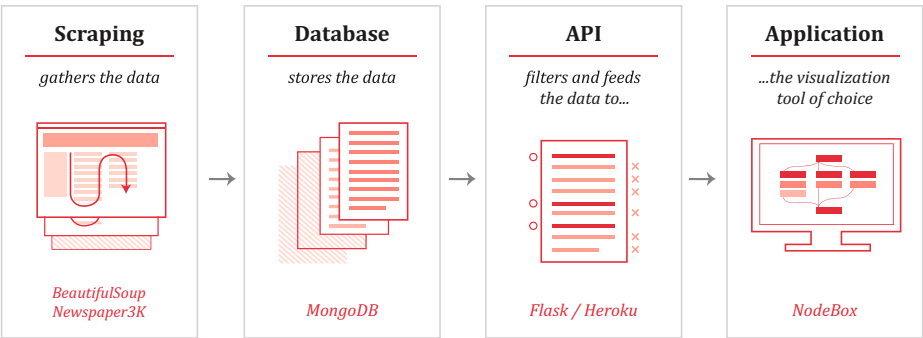
**Fig. 3** The *time-sets* visualization of news articles related to “immigration”, “migration”, and “refuge” topics published in 2015

are then placed consecutively on a horizontal axis to allow comparisons of the individual months side-by-side.

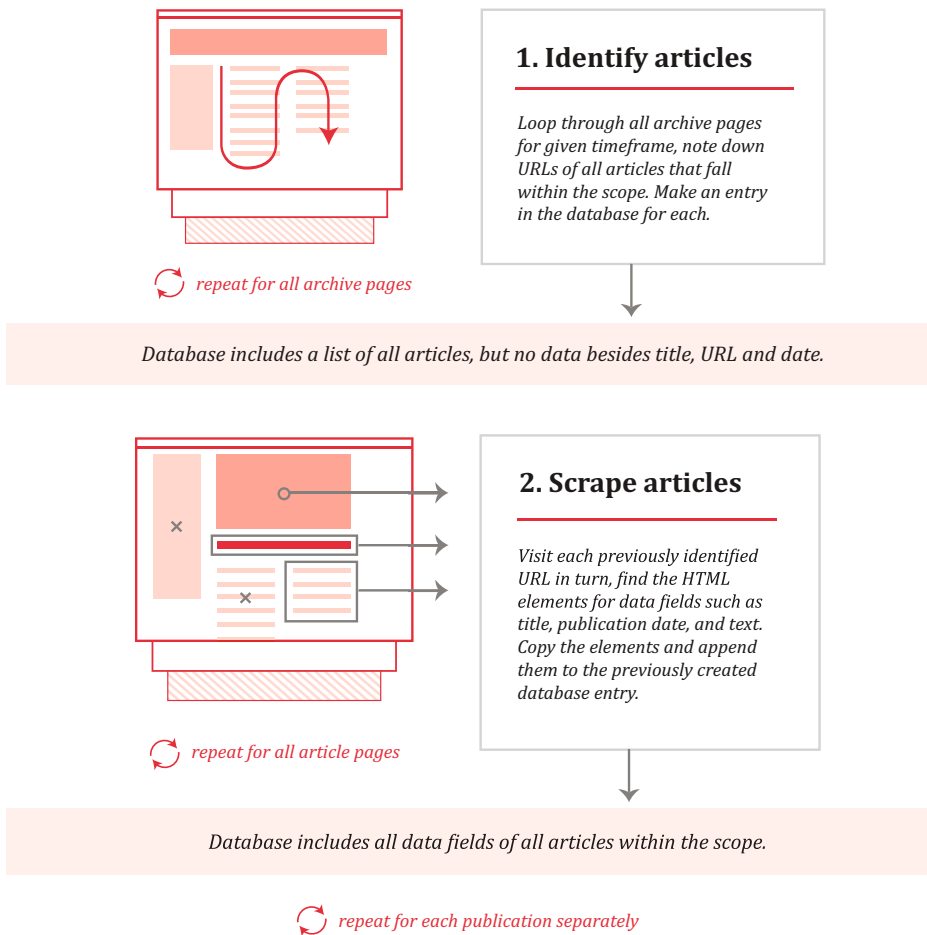
The prototype extends the basic form of the *time-sets* visualization described earlier by placing horizontally aligned linear diagrams on a vertical axis, to allow comparisons in a 2D space. As can be seen, the horizontal representations are created for each online news source (i.e., publications A, B, and C) individually, resulting in a *time-sets* visualization supporting comparisons of each news source, not only month-by-month horizontally, but also different news sources against each other vertically.

4.1 Implementation

We have used a modular process for creating the dataset and accessing it through the *time-sets* visualization prototype. Figure 4 shows the different components of this process pipeline.



**Fig. 4** The process pipeline of *time-sets* prototype



**Fig. 5** The scraping pipeline for creating the articles dataset

The primary method of data acquisition for our test dataset was through “scraping” – i.e., using automated scripts crawling a series of online news archive pages<sup>9</sup>. The scripts we used for scraping were written in Python, and utilized the BeautifulSoup<sup>10</sup> and Newspaper3K<sup>11</sup> Python libraries to identify and extract articles, in conjunction with the PyMongo<sup>12</sup> tools, to pass the scraped data to the MongoDB<sup>13</sup> database.

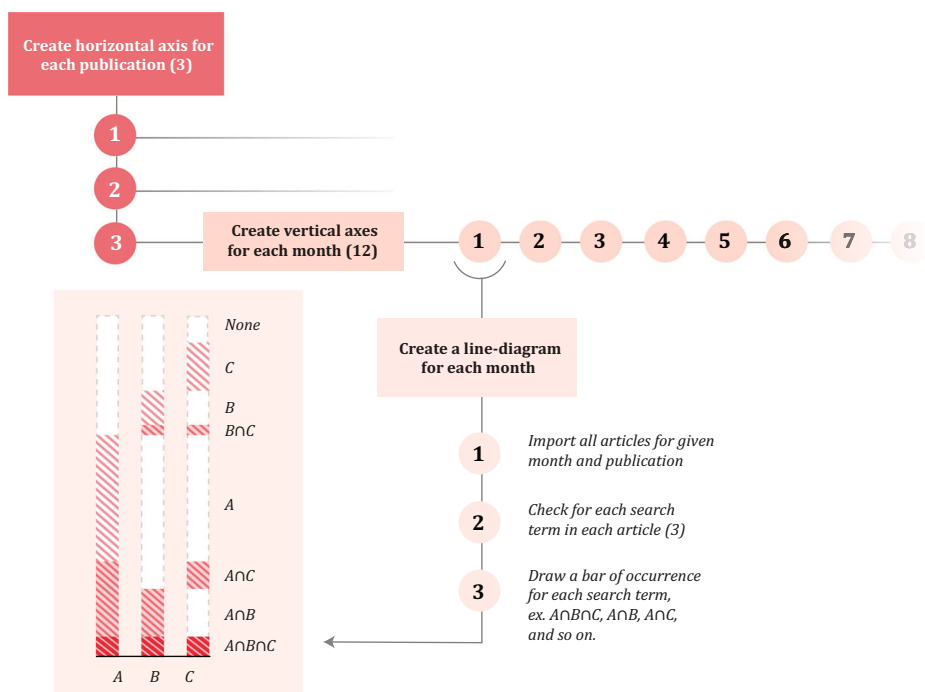
<sup>9</sup>The European Commission Directive on Copyright in the Digital Single Market, Proposal 2016 (<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016PC0593&from=en>), provides “wider opportunities to use copyrighted material for education, research, cultural heritage and disability (through so-called ‘exceptions’),” and allowances for automated data gathering methods to these ends. Also see, European Commission Press Release, September 2018 ([http://europa.eu/rapid/press-release\\_STATEMENT-18-5761\\_en.htm](http://europa.eu/rapid/press-release_STATEMENT-18-5761_en.htm)).

<sup>10</sup><https://www.crummy.com/software/BeautifulSoup/>

<sup>11</sup><http://newspaper.readthedocs.io/>

<sup>12</sup><https://api.mongodb.com/python/current/>

<sup>13</sup><https://www.mongodb.com/>



**Fig. 6** The visualization pipeline of the *time-sets* prototype

The scripts worked in two consecutive loops (see Fig. 5), first using BeautifulSoup to iteratively gather a list of URLs to be scraped from the online news archive pages, and then on a second pass to scan each page and use Newspaper3K to copy the article elements to the database. The scripts we used for each publication source varied in turn, to account for differences in the web page structure, as well as to excluded dates and topics not included within the predefined scope (as described earlier).

The NoSQL<sup>14</sup> MongoDB database was used to store the scraped articles in two separate collections. The raw scraped data was initially housed in one collection, from which a second “streamlined” version was generated using Python natural language processing tools – to remove stop words, count occurrences of certain predefined keywords, stem words<sup>15</sup>, and build a short-list of search terms which were deemed interesting for the purposes of visualization.

A Flask<sup>16</sup> Python microframework application was then deployed on a free Heroku<sup>17</sup> instance to serve up the API used to generate the visualization. In short, the Flask application pulled article data from the database and returned the data in a plain text format, which in this case included the number of word occurrences and co-occurrences within a given timeframe. By altering the API, this step could also provide occurrences and co-occurrences of video or audio clips in articles, number of images per article, article length, or any number

<sup>14</sup><https://en.wikipedia.org/wiki/NoSQL>

<sup>15</sup><https://snowballstem.org/>

<sup>16</sup><http://flask.pocoo.org/>

<sup>17</sup><https://www.heroku.com/>

of other multimedia content characteristics, or metadata variables that might conceivably be visualized.

With a functional API made available through the above process, the visualization itself could then be built using any number of existing visualization toolkits or libraries. We decided to implement our *time-sets* visualization prototype using NodeBox Live<sup>18</sup>, a node-based visualization platform that can be run in a web browser. Figure 6 shows a schematic view of the visualization pipeline, which is comprised of several distinct steps:

1. creating a horizontal axis for each online news source (publication),
2. creating 12 vertical axes (one per each month of the year) for each horizontal news source,
3. pulling the relevant data from the dataset for each particular month,
4. checking each article for each of the search keywords (words or terms),
5. compiling a line diagram for that month.

We have tested our prototype with a dataset of several hundred thousand articles. While the current prototype is very functional, its efficiency and scalability could be further improved using a more robust set of tools than those provided by NodeBox Live.

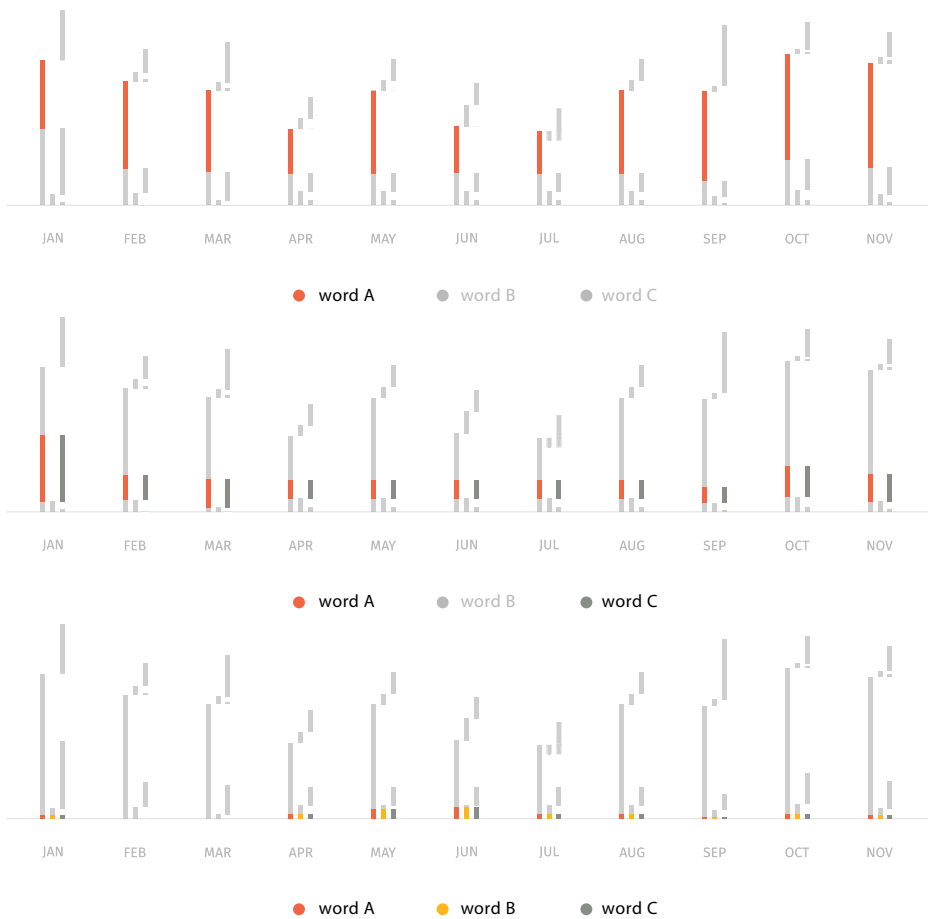
## 4.2 Interactivity with multimedia content

Although static visualizations generated using the *time-sets* prototype support visual analysis of co-occurrences over several publications across time, addition of interactive elements would extend its capabilities, particularly in allowing users to further investigate in a detailed view mode interesting patterns discovered in an overview mode. Furthermore, interactive elements would also support access to the underlying multimedia data (e.g., images, video, etc.) which enrich most online publications. Similarly, the data fed into the visualization could, for instance, be a live stream of media coverage, word use, incidence of video elements in coverage, image use (original photographs vs. stock photo-bank images), image content-based analysis, co-occurrence of web/print/radio coverage of news items, occurrence of reader comments, social media shares over a certain threshold, or any number of other multimedia variables. Integrating interactive elements into the user interface of the prototype could also further extend the applications of the multimedia data elements of the visualization. Therefore, in this section we present several interactive elements we have investigated so far to enhance the capabilities of the current prototype.

A key interactive element that allows further user insight into the dataset is *brushing* – i.e., visually separating and highlighting selected subsets of the visualization for individual scrutiny by, for instance, graying out the non-selected subsets. Brushing is supported by *time-sets* visualization in two different manners. Figure 7 shows an example of what could be termed as *in-place brushing*, where highlighted elements remain visually in the context of grayed-out elements without visually moving them in the 2D space. In this example, it is easy to see the number of articles in each month in which only *Word A* occurs (top), or both words *A* and *C* occur (middle), or all three words *A*, *B* and *C* occur (bottom).

Although *in-place brushing* makes it possible to visually isolate elements of interest, those elements do not necessarily share a common horizontal axis (e.g., see Fig. 7 middle).

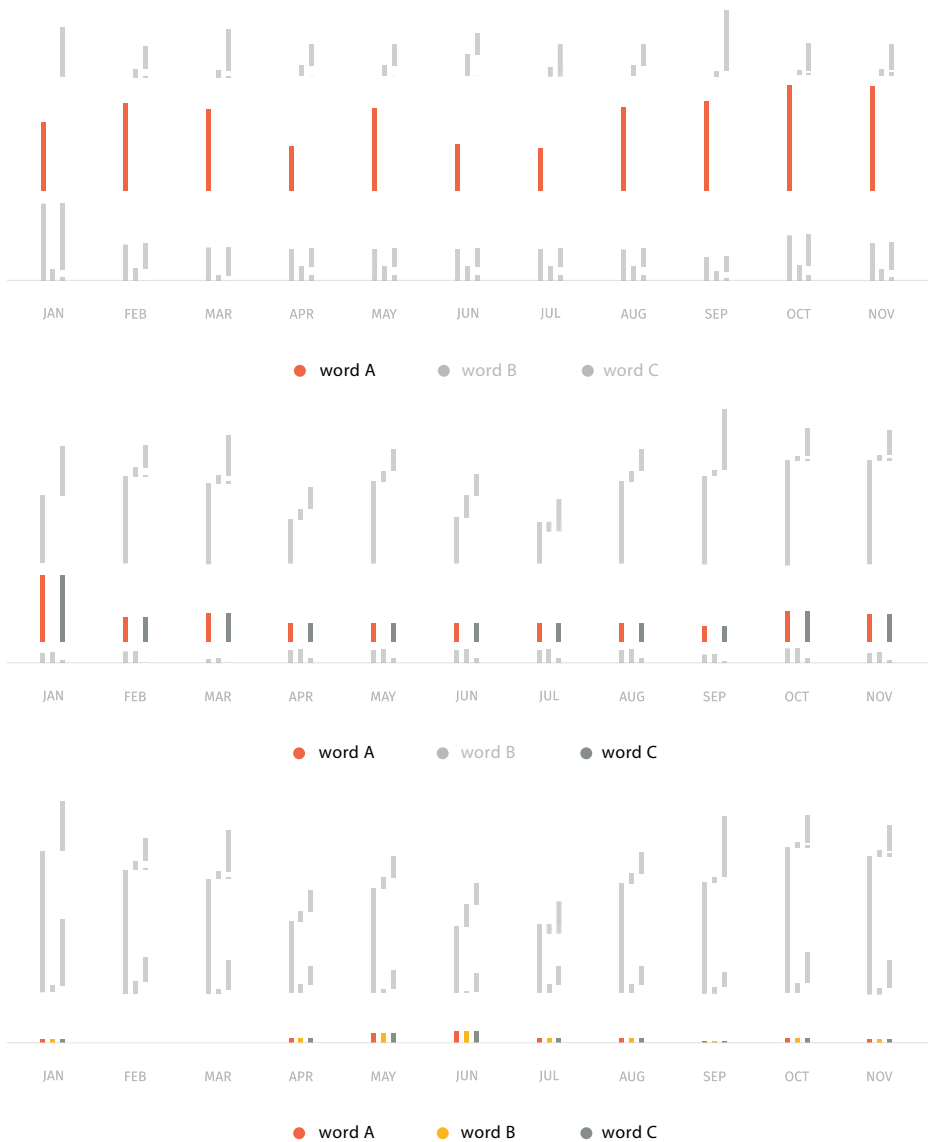
<sup>18</sup><https://nodebox.live/>



**Fig. 7** An example of *in-place brushing*, showing articles with only *Word A* (top), words *A* and *C* (middle), and words *A*, *B* and *C* (bottom)

As mentioned earlier, tasks which require visual positioning along non-aligned axes are more difficult than those carried out on a common axis [9] (e.g., see Fig. 7 bottom). Therefore, in *time-sets* visualization we also propose an alternative brushing, which we refer to as *common-axis brushing*. Figure 8 shows an example of *common-axis brushing*, in which highlighted elements are moved and isolated in 2D space, in relation to grayed-out elements, to align them visually on a shared horizontal axis.

Another interactive method which supports detailed inspection of selected elements in the context of non-selected elements (i.e., focus-in-context) is *bifocal zoom* [20]. Figure 9 shows an example of *bifocal zoom*, in which the co-occurrence data for the selected month of June is shown in detail within the context of other non-selected months. In this example, the non-selected months have been grayed-out in addition to being shown in less detail. The other alternative would be to only show them in less detail, without grayed-out. This, however, may lead to the viewer making misleading comparisons of the number of articles



**Fig. 8** An example of *common-axis brushing*, showing articles with only *Word A* (top), words *A* and *C* (middle), and words *A*, *B* and *C* (bottom)

shown in the zoomed and non-zoomed parts of the visualization (e.g., wrong comparison of the number of articles in weeks versus months).

It is also possible to use *time-sets* with other conventional visual analysis tools used in journalism. Figure 10 shows an example of the *time-sets* visualization, along with a bar chart and line chart, which show the total number of articles published each month containing the

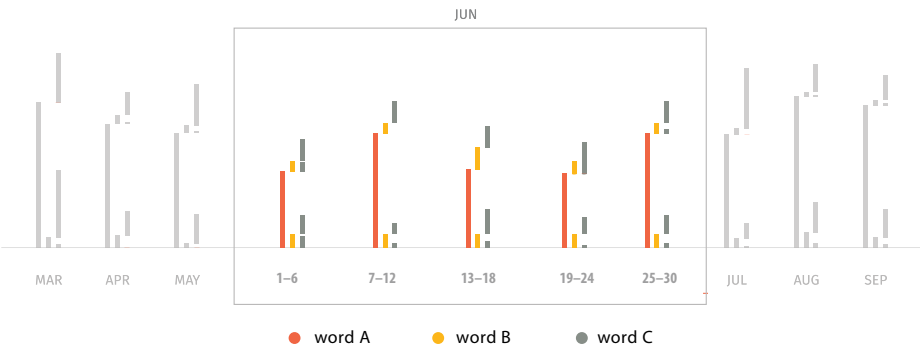


Fig. 9 An example of the *bifocal zoom*, showing detailed view of a selected month (June) in the context of other non-selected months

three selected words. In the case of being used in combination with other visualizations, it is important to support *linking* between *time-sets* and those visualizations, in addition to the other interactive elements described above.

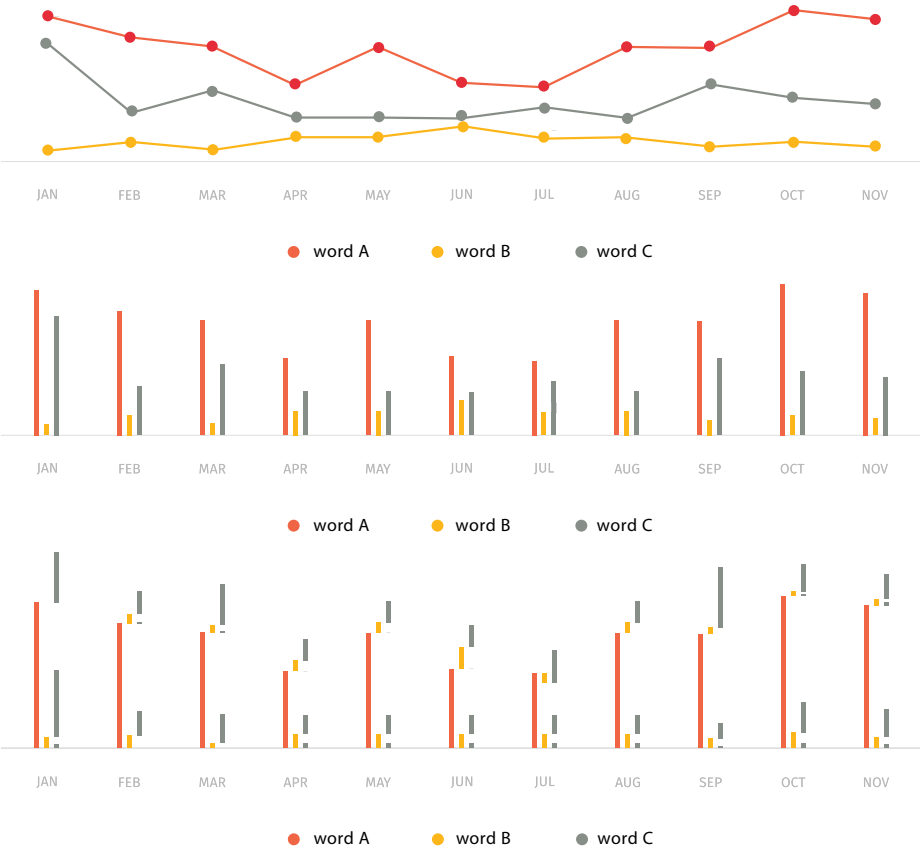
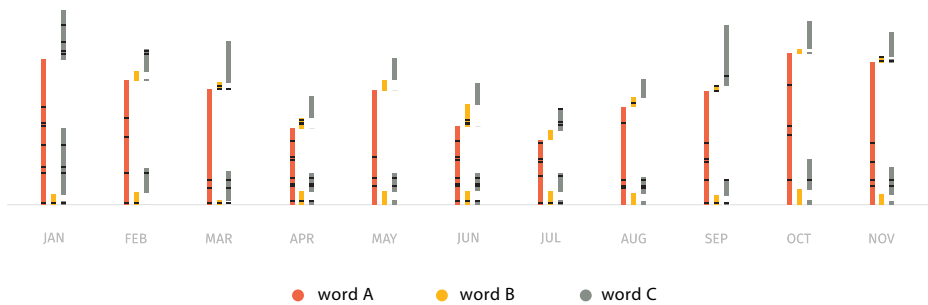


Fig. 10 An example of the *time-sets* visualization (bottom), along with a bar chart (middle) and line chart (top) showing the total number of articles published containing each of the selected words





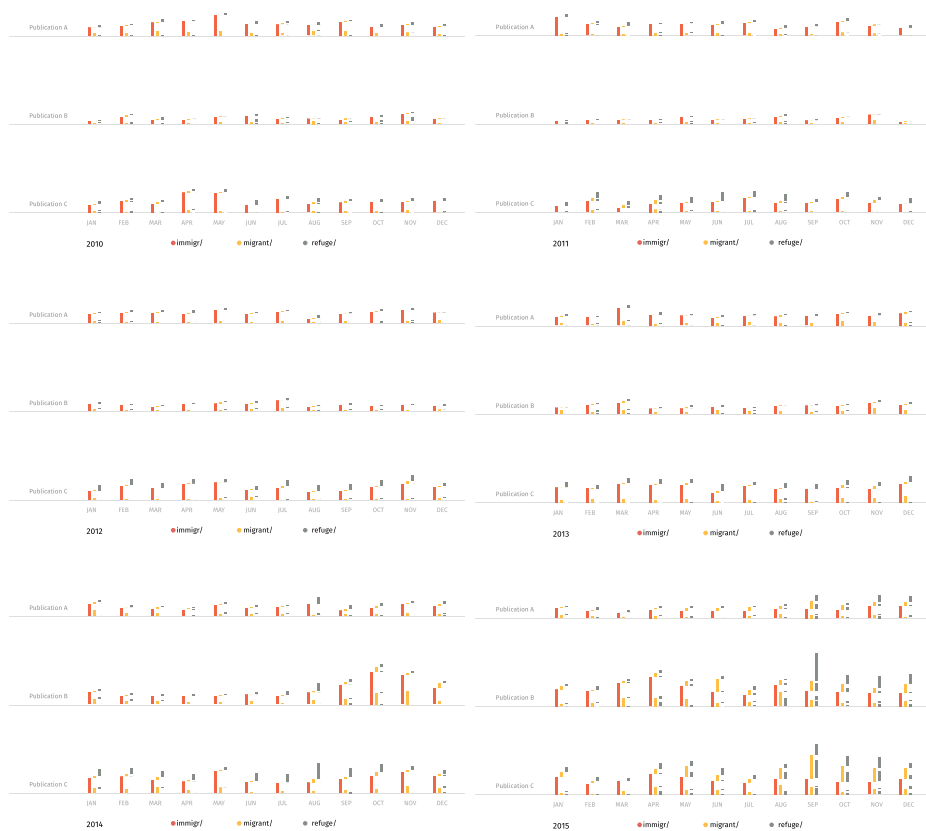
**Fig. 11** An example of the *time-sets* visualization in which articles containing video clips have been *tick-marked*

Finally, in addition to allowing users to view co-occurrences of visualized data, *time-sets* could be used to access details of individual articles contained in the dataset being visualized, as well as the articles themselves, through links to the actual online material. The articles could be ordered chronologically (or according to some other predefined criteria) vertically on each of the line segments of *time-sets*. Being able to examine individual articles in this manner would further improve the transparency of the dataset, and enable the user to locate specific events on the timeline that may have had an effect on the overall narrative. This could also utilize a *mouse over* type of interaction, bringing forth for instance a small preview of the article title, its publication metadata, and headline image. The visualization also supports the use of *tick-marks* to indicate all the articles within the dataset that included videos, stock images, original event photos, reader comments, social media shares, and so on. Figure 11 shows an example of this concept, where articles containing video clips have been marked on each of the line segments of *time-sets*. The main limitation of this idea in practice is that if there are a large number of articles in the dataset then the *tick-marks* will be difficult to distinguish from each other. In such cases, other zooming techniques could be provided (e.g., vertical bifocal zoom) to allow detailed inspection of individual line segments.

## 5 Illustrative use case examples

Although a full discussion of all the kinds of findings one could obtain through visual analysis of the visualizations generated using our prototype is not necessary here, we provide a few illustrative examples related to our test dataset to demonstrate the potential effectiveness of the *time-sets* visualization.

Figure 12 shows a series of visualizations generated for years 2010 (top left) to 2015 (bottom right). In this example it is obviously not possible to observe any specific details, but rather this example shows the potential of *time-sets* when used as *small multiples* [35]. For instance, from these small multiples it can be observed that *Publication B* (middle row in each year) does not publish many articles related to our chosen topics between 2010 and 2013, but suddenly seems to increase its related articles from the end of 2014 to 2015. This is particularly observable in the case of articles related to “immigration” (shown in red) initially, and then “refuge” (shown in gray) towards the end of 2015. In comparison, for the

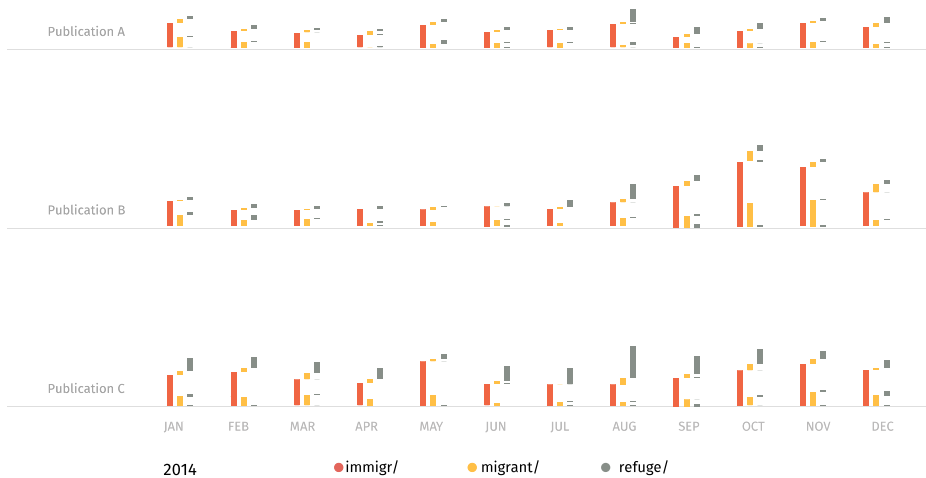


**Fig. 12** The *time-sets* visualizations of news articles related to “immigration”, “migration”, and “refuge” topics published between 2010 (top left) and 2015 (bottom right), shown as *small multiples*

same period (2010–2015) *Publication A* (top row in each year) remains largely unchanged. Similarly, one can observe that *Publication C* (bottom row in each year) generally publishes more articles related to “immigration”, and to some extent “migration” and “refuge”, than the other two publications.

More specific observations could also be made by visual analysis of individual *time-sets* visualizations, or by comparing two or more of them closely together. For example, looking at the visualization for 2014 (Fig. 13), one can observe that *Publication C* consistently published article related to the “refuge” topic (shown in gray) on its own more than the other two publications. Similarly, the number of articles related to both “immigration” (shown in red) and “migration” (shown in yellow) remain more consistent in *Publication C*. This may reflect the fact that perhaps *Publication C* does not change the topics of its articles to suit the current trends and issues. On the other hand, *Publication B* seems to follow current trends more closely (i.e., “hot topics”), as can be seen from a comparison of Figs. 3 and 13.

Finally, comparing the visualizations for 2014 (Fig. 13) and 2015 (Fig. 3), one can see that there is a substantial increase in 2015 in the use of terms related to “migration” and “refuge”, either on their own or in combination together, particularly in publications *B* and



**Fig. 13** The *time-sets* visualization of news articles related to “immigration”, “migration”, and “refugee” topics published in 2014

*C*, but also to some extent in *Publication A*. Although this is likely to have been caused by a large number of refugees and migrants arriving in Europe in the summer of 2015, nevertheless, it may also show an increasing shift in the use of terms related “migration”, rather than those related to “immigration”.

## 5.1 Use cases with multimedia content

In the above examples we have only discussed use cases focusing on the analysis of co-occurrences of terms in textual content. This has primarily been because of the fact that – due to a lack of access to large storage and computational resources required – our current dataset only contains textual material and metadata, as discussed in Section 4.

However, as mentioned earlier, these days most online news publications contain not only text, but also other multimedia content such as video and visual material. Clearly the analysis of the co-occurrences of these types of multimedia content – either on their own, or in combination with textual content – is becoming increasingly important in journalism.

Our previous exploratory work (for a discussion, see [19]) has, for instance, shown that an online image search for the term “immigrant” resulted in shockingly different types of images being reported in different European countries. For example, while in some countries immigrants were portrayed as an encroaching threat of unwashed masses, in others, they were shown as happy contributing members of society.

With more computational and storage resources available, it would be feasible to create datasets of published news articles which include a greater range of multimedia content. With such datasets, it would then be possible to use the *time-sets* visualization to undertake a range of visual analysis of publication patterns, such as various forms of co-occurrences across publications and time periods.

As shown in Fig. 11, *time-sets* supports the use of *tick-marks* to indicate which articles include what types of multimedia content (e.g., videos, stock images, original event photos, etc.). An example use case of this tool might be to investigate what types of video footage are

included in news articles that contain the term “migration”. Depending on the patterns being observed, one could then perhaps draw conclusions about the impartiality or verifiability of news items, for instance, based on the inclusion of verified video footage – as opposed to inclusion of unverified videos, which might be due to a less-than-accurate or even a totally biased agenda on the part of the publication.

Similarly, with further analysis of the content of images, videos, audio, and other multimedia material their metadata in the dataset could be enhanced, and then visualized using *time-sets*. This would allow, for example, identifying patterns in the content of images being included – i.e., co-occurring – in news article relating to a particular term (e.g., the types of images being used along with the term “immigrant”, as pointed out above).

## 6 User study

We have conducted a user study to evaluate the effectiveness of *time-sets* visualization for visual analysis tasks similar to those described in the previous illustrative use case example. Although we envisage this type of visualization to be used primarily for journalistic purposes, the types of visual analyses described above should be achievable for any user, with or without professional training in journalism or media analysis. As such, our user study targeted ordinary users and aimed to evaluate the visual effectiveness of *time-sets* without direct reference to any specific journalistic activity.

### 6.1 Methodology

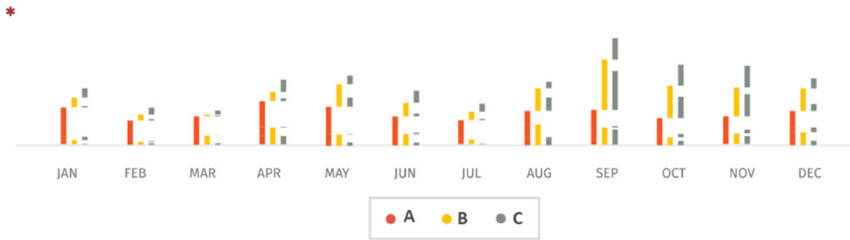
Our user study was created and conducted using the LimeSurvey<sup>19</sup> online survey tool. The study started with a brief introductory statement of intent and confirmation of the participant’s consent. The study participants were then presented with a short description of the *time-sets* visualization, followed by three training questions similar to those used in the actual study task questions (see below).

Each multiple-choice training question used the same single-axis *time-sets* visualization depicting the occurrence of three word variables in a single publication over twelve months (see Fig. 14a). The data used in the sample visualizations was anonymized, and presented in the form publications 1, 2, 3 and words A, B and C, to avoid any unintended biases associated with certain publications or topics. The training questions were interactive and automatically corrected the participant’s answers with a visual aid and explanation of the correct answer (see Fig. 14b). These initial three questions were designed to train the participants to: a) identify a single variable in a publication, b) compare the level of variable occurrences along a single axis, and c) identify co-occurrences of variables. Following each training question, the participants were asked to rate its difficulty level (as with the actual study tasks to follow).

After the completion of the three training questions, the participants were presented with the study task questions, each of which was followed by a difficulty rating questions. Participants were informed that their answers to task questions would no longer be corrected (as was done with the training questions), and that their responses to the task questions would be timed. However, the responses to the difficulty rating questions were not timed.

---

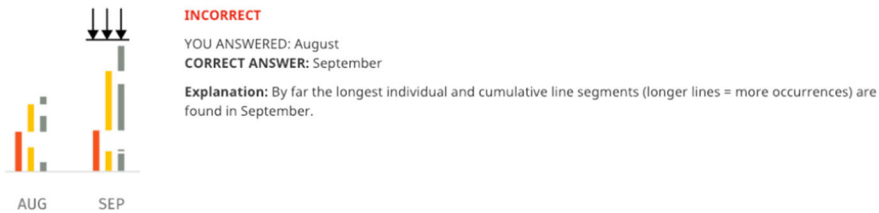
<sup>19</sup><https://www.limesurvey.org/>



This time-sets diagram represents all the articles by a certain publication in a single year that contain keywords A, B and/or C. In which month were the **most** articles published containing some or all of these words?

- ☐ August
- ☐ September
- ☐ October

a



b

**Fig. 14** Example of a training question **a**, and its feedback **b**, as shown in LimeSurvey

## 6.2 Task Questions

Table 1 presents the six task questions used in this study. Each of the questions was presented to the participants individually along with a single *time-sets* visualization image. Figure 15 shows the two images used, where image Fig. 15a was used for questions 1–3, and image Fig. 15b was used for questions 4–6. As such, questions 1–3 required making comparisons or identifying word co-occurrences across a single publication, while questions 4–6 required making comparisons or identifying word co-occurrences across three publications. Table 2 provides a summary of all the study task question types.

## 6.3 Participants

The participants were convenience sampled from both personal and professional circles of acquaintance, as well as the extended social media community. In total, 65 people took part in the online study, of which 25 did not proceed beyond the training questions or did not fully complete all the study tasks questions, and were therefore excluded from the data analysis. Responses from 2 other participants were also excluded from further analysis due to very long pauses (10–15 minutes) during one of the task questions. The remaining completed responses from 38 participants were subsequently analyzed.

**Table 1** The study task questions, with answer options shown in brackets, and correct answers underlined

No.	Question	Image
1.	How did the occurrence of word C change from September to December? ( <i>Increased, <u>Decreased</u>, Stayed mostly the same</i> )	Figure 15a
2.	How did the co-occurrence of all three words change from August to September? ( <i>Increased, <u>Decreased</u>, Stayed mostly the same</i> )	Figure 15a
3.	In which of the following was there no apparent co-occurrence of words A and C? ( <i>January, <u>July</u>, October, None of the above</i> )	Figure 15a
4.	In which of these months did Publication 1 use all three words together more than Publication 3? ( <i>October, November, December, <u>None of the above</u></i> )	Figure 15b
5.	Which of the three publications generally used word C least often the entire year? ( <i><u>Publication 1</u>, Publication 2, Publication 3</i> )	Figure 15b
6.	Which co-occurrence of two words increased the most from August to September in all three publications? ( <i>A and B, A and C, <u>B and C</u></i> )	Figure 15b



**Fig. 15** The two images used for the study task questions 1–3 **a** and 4–6 **b**

The participants' age distribution was 34.2% (13) aged 20–30, 31.6% (12) aged 31–40, 18.4% (7) aged 41–50, 7.9% (3) aged 51–60, and 7.9% (3) above 60. Self-reported professions included 26.3% (10) academics, 13.2% (5) students, 5.3% (2) journalists, with the remaining 55.3% (21) choosing “Other” as their profession (including several designers, a botanist, a lawyer, a public health official, and a government analyst). Most of the participants (28 Yes, 9 No, 1 N/A) had 20/20 or equivalent corrected vision, and almost none (0 Yes, 37 No, 1 N/A) had impaired color vision (protanopia, deuteranopia or similar).

**Table 2** Summary of the study task question types

Question	Publications	Visual element	Visual task	Image
1.	single	occurrence	comparison across time	Figure 15a
2.	single	co-occurrence	comparison across time	Figure 15a
3.	single	co-occurrence	identification	Figure 15a
4.	three	co-occurrence	identification	Figure 15b
5.	three	occurrence	comparison across time	Figure 15b
6.	three	co-occurrence	comparison across time	Figure 15b

**Table 3** Summary of the results of the study for each task question

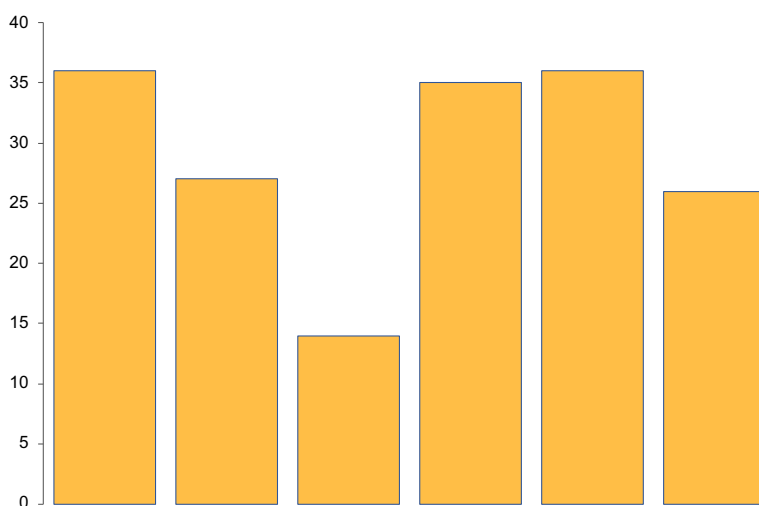
Question	Accuracy ( <i>n</i> =38)		Time to complete ( <i>seconds</i> )		Difficulty rating (1: <i>v. easy</i> , 7: <i>v. difficult</i> )	
	No.	%	Mean	SD	Mean	SD
1.	36	94.74	38.14	17.75	2.29	1.11
2.	27	71.05	35.88	19.53	2.47	1.25
3.	14	36.84	46.04	19.61	3.00	1.29
4.	35	92.11	67.66	41.12	3.87	1.38
5.	36	94.74	40.83	30.37	2.58	1.13
6.	26	68.42	64.58	37.18	3.97	1.33

## 6.4 Results

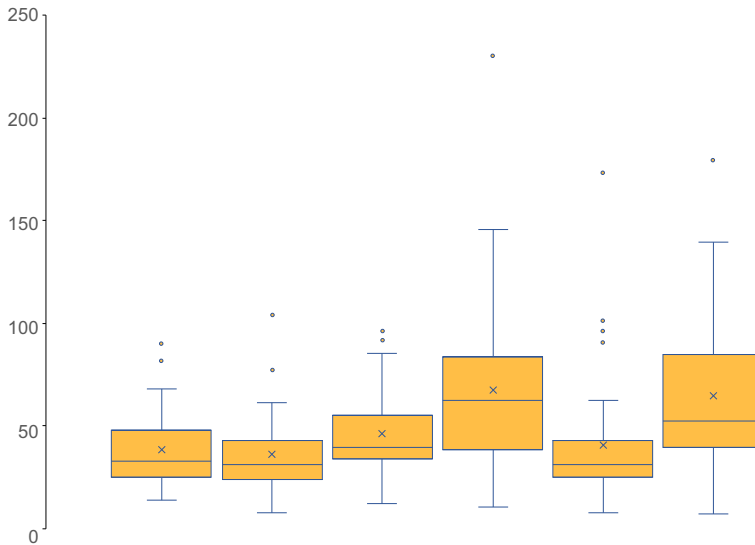
For each of the study task questions, we analyzed the accuracy of the participants' responses, the time taken to answer them, and the difficulty ratings given to them by the participants. Table 3 presents a summary of the results of these analyses.

Figure 16 shows a bar chart of the number of correct answers given to each of the study task questions by the participants ( $N = 38$ ). We had initially envisaged that questions 1–3, which relied on a single publication visualization, would be easier than questions 4–6, and would therefore have a higher response accuracy. This, however, was not the case. In fact Question 3 had the lowest accuracy rate amongst all the 6 questions. While the correct answer for this question was “July”, 57.89% (22) of the participants answered “None of the above” – thus failing to see the co-occurrences of words *A* and *C* in the visualization shown in Fig. 15a.

The accuracy of responses to Question 2 (with the correct answer of “Increased”) was also lower than those of most questions, but in this case the number of incorrect answers

**Fig. 16** Number of correct answers given to each of the study task questions 1–6, ( $N = 38$ )





**Fig. 17** Time taken to complete each of the study task questions 1–6, (*time in seconds*)

was almost equally divided between the two other options (“Decreased” and “Stayed mostly the same”).

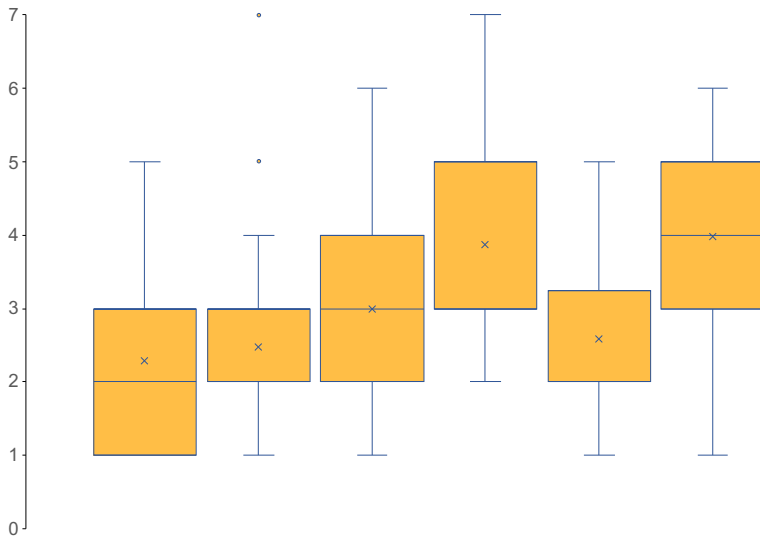
In terms of the questions 4–6, which relied on the use of a visualization with three publications (see Fig. 15b), we had envisaged that Question 6 will be the most difficult one to answer, with the lowest accuracy rate. This almost proved to be the case, with Question 6 (with the correct answer of “B and C”) having the second lowest accuracy rate of 68.42% (26 out of 38). Most of the incorrect answers (26.32%, or 10 out of 38) for Question 6 were for “A and C”. The other two questions in this category (4 and 5) had surprisingly high accuracy rates (92.11% and 94.74%, respectively).

Figure 17 shows a box-and-whisker plot for the time taken (in seconds) to answer each of the study task questions by the participants. As expected, questions relying on the visualization of a single publication at a time (questions 1–3, and 4 to some extent) took less time to answer. It is interesting to note that while Question 3 was in this category, and did not take much longer than the other questions to answer, it resulted in much fewer correct answers. This may indicate that either the participants misinterpreted the question, or the visualization used with this question was somehow misleading – i.e., the participants quickly answered the question thinking that their answer was correct.

Finally, Fig. 18 shows a box-and-whisker plot for the difficulty ratings given (with anchors, 1: very easy, 7: very difficult) to each of the study task questions by the participants. The results of this analysis largely corroborate the results of the previous analysis in time taken to answer each of the study task questions. As can be seen, questions which took longer to answer were also rated as being more difficult. Overall, most of the questions were rated as having low to medium level of difficulty.

## 6.5 Discussion

Based on the results reported above, our study has shown that the *time-sets* visualization could be effective in supporting users in making visual comparisons of co-occurrences



**Fig. 18** Difficulty ratings given to each of the study task questions 1–6, (anchors, 1: very easy, 7: very difficult)

across a large number of articles in several publications over a long period of time. In most cases, our study participants were able to complete the study task questions with a high level of accuracy, in a reasonably short amount of time, and without much perceived difficulty.

However, the study also highlighted the fact that in some cases (e.g., Question 3) the visualization may be misinterpreted by some users. It seems that this might be the case when some line segments are very short in comparison to others (i.e., the occurrence or co-occurrence of their related variables are much smaller than others), making them difficult to see at the normal visualization zoom level. Although, the use of zooming tools, such as those described earlier, could perhaps solve this problem, it might also be necessary to set minimum levels for the occurrence or co-occurrence values that are considered significant enough to be displayed.

It is also important to note that in this study we only relied on the use of static visualizations generated using our *time-sets* interactive prototype tool. While this provided our study participants with the same – and therefore comparable – visualization images for each study task, in real-life settings the users are much more likely to want to interact with the visualization prototype itself, as is the case in most visual analytic type tasks. Therefore, it is still necessary to conduct a further evaluation of the prototype itself, once all its interactive tools have been fully implemented.

Similarly, as mentioned earlier, due to computational and storage limitations, the dataset we have created only contains textual news material and metadata. As such, our current study did not include any tasks specifically dealing with various types of multimedia content. While we envisage that the visualizations of co-occurrences, as provided by *time-sets*, will remain effective regardless of the underlying data, any future evaluations of the interactive prototype should also test its capabilities with other multimedia news content such as audio, video or static images.

Finally, we would like to point out that in the study reported here only two of the participants (out of a total of 38) were actually journalists. Although this study investigated

the effectiveness of *time-sets* visualization in supporting visual detection of co-occurrence patterns – which can be carried out by anyone with normal vision – any evaluation of the journalism-related activities using an interactive version of the *time-sets* prototype should be done with users who are specifically interested in, and have the necessary skills to perform, such journalistic tasks.

## 7 Conclusions

In this paper, we have argued the need for visualizations that support tracking changes – particularly those related to the co-occurrence of variables – in online media narrative patterns over time and across publications. We have also discussed why temporal comparisons of co-occurrence choices in media coverage can be more useful than simply counting their occurrences individually.

The potential of *time-sets* visualization, which aims to support temporal comparisons of such co-occurrences, has been investigated through development of an interactive prototype. We have provided examples of visualizations generated using this prototype on a large dataset of news articles scraped from three online publications, spanning a period of six years (2010–2015). These examples show that indeed interesting co-occurrence and other temporal patterns can be detected from such *time-sets* visualizations.

We have also carried out a user study to evaluate the effectiveness of *time-sets* visualization in supporting visual analysis tasks similar to those described in our use case example. The results of this study have shown that such visual analyses can be completed by ordinary users, who do not necessarily have any professional training in journalism.

Although our current prototype has a limited number of interactive tools, these could be extended easily along the lines of interactive functionality that we have identified. We are also aiming to investigate a number of alternatives to the *time-sets* visualization, including the use of space-filling mosaics [23], which have been shown to be more effective than linear diagrams for some tasks involving visual comparisons of set relationships [24], but without involving comparisons across multiple axes or time.

**Acknowledgements** We would like to gratefully acknowledge the contributions of everyone who anonymously participated in our online user study.

**Funding Information** Open access funding provided by Aalto University.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Aigner W, Miksch S, Schumann H, Tominski C (2011) Visualization of Time-Oriented data. Human-Computer interaction series. Springer, London
2. Alsallakh B, Micallef L, Aigner W, Hauser H, Miksch S, Rodgers P (2016) The state-of-the-art of set visualization. *Comput Graph Forum* 35(1):234–260. <https://doi.org/10.1111/cgf.12722>
3. Angus D, Smith A, Wiles J (2012) Conceptual recurrence plots: Revealing patterns in human discourse. *IEEE Trans Vis Comput Graph* 18(6):988–997. <https://doi.org/10.1109/TVCG.2011.100>
4. Angus D, Watson B, Smith A, Gallois C, Wiles J (2012) Visualising conversation structure across time: Insights into effective doctor-patient consultations. *PLOS One* 7(6):1–12. <https://doi.org/10.1371/journal.pone.0038014>

5. Bertin J (1981) Graphics and graphic information-processing. de Gruyter, Berlin
6. Bullinaria JA, Levy JP (2007) Extracting semantic representations from word co-occurrence statistics: a computational study. *Behav Res Methods* 39(3):510–526. <https://doi.org/10.3758/BF03193020>
7. Chapman P, Stapleton G, Rodgers P, Micallef L, Blake A (2014) Visualizing sets: An empirical comparison of diagram types. In: Dwyer T, Purchase H, Delaney A (eds) *Diagrammatic Representation and Inference, Proceedings of the International Conference on Theory and Application of Diagrams*. Springer, Berlin, pp 146–160. [https://doi.org/10.1007/978-3-662-44043-8\\_18](https://doi.org/10.1007/978-3-662-44043-8_18)
8. Choi J (2009) Diversity in foreign news in US newspapers before and after the invasion of Iraq. *Int Commun Gaz* 71(6):525–542. <https://doi.org/10.1177/1748048509339788>
9. Cleveland W, McGill R (1985) Graphical perception and graphical methods for analyzing scientific data. *Science* 229(4716):828–833
10. Fahmy S, Emad MA (2011) Al-Jazeera vs Al-Jazeera: a comparison of the network's English and Arabic online coverage of the US/al Qaeda conflict. *Int Commun Gaz* 73(3):216–232. <https://doi.org/10.1177/1748048510393656>
11. Friendly M (2008) The golden age of statistical graphics. *Stat Sci* 23(4):502–535. <https://doi.org/10.1214/08-STS268>
12. Gan F, Teo JL, Detenber BH (2005) Framing the battle for the White House: a comparison of two national newspapers' coverage of the 2000 United States presidential election. *Int Commun Gaz* 67(5):441–467. <https://doi.org/10.1177/0016549205056052>
13. Gottfried B (2014) Set space diagrams. *J Visu Lang Comput* 25(4):518–532. <https://doi.org/10.1016/j.jvlc.2014.04.003>
14. Gottfried B (2015) A comparative study of linear and region based diagrams. *J Spatial Inf Sci* 10:3–20. <https://doi.org/10.5311/JOSIS.2015.10.187>
15. Huang Y, Fahmy S (2011) Same events, two stories: Comparing the photographic coverage of the 2008 anti-China/Olympics demonstrations in Chinese and US newspapers. *Int Commun Gaz* 73(8):732–752. <https://doi.org/10.1177/1748048511420091>
16. Humphries B, Radice M, Lauzier S (2017) Comparing “insider” and “outsider” news coverage of the 2014 ebola outbreak. *Can J Publ Health* 108(4):381–387. <https://doi.org/10.17269/cjph.108.5904>
17. Jensen KB (ed) (2012) *A Handbook of Media and Communication Research: Qualitative and Quantitative Methodologies*, 2nd edn. Routledge, New York
18. Kim D, Johnson TJ (2009) A shift in media credibility: Comparing internet and traditional news sources in South Korea. *Int Commun Gaz* 71(4):283–302. <https://doi.org/10.1177/1748048509102182>
19. Koivunen L (2018) Narrative variance: Visualizing patterns in media coverage. Master's thesis, Aalto University
20. Leung YK, Apperley MD (1994) A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans Comput-Hum Interact* 1(2):126–160. <https://doi.org/10.1145/180171.180173>
21. Lex A, Gehlenborg N, Strobel H, Vuilleumot R, Pfister H (2014) UpSet: visualization of intersecting sets. *IEEE Trans Vis Comput Graph* 20(12):1983–1992. <https://doi.org/10.1109/TVCG.2014.2346248>
22. Lim J (2018) Representation of data journalism practices in the South Korean and US television news. *International Communication Gazette*. <https://doi.org/10.1177/1748048518759194>
23. Luz S, Masoodian M (2011) Comparing static gantt and mosaic charts for visualization of task schedules. In: *Proceeding of the 15th international conference information visualisation*, pp 182–187. <https://doi.org/10.1109/IV.2011.53>
24. Luz S, Masoodian M (2018) A comparison of linear and mosaic diagrams for set visualization. *Information Visualization*. <https://doi.org/10.1177/1473871618754343>
25. Luz S, Sheehan S (2014) A graph based abstraction of textual concordances and two renderings for their interactive visualisation. In: *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, AVI '14*. ACM, pp 293–296. <https://doi.org/10.1145/2598153.2598187>
26. Mahony I (2010) Diverging frames: A comparison of Indonesian and Australian press portrayals of terrorism and Islamic groups in Indonesia. *Int Commun Gaz* 72(8):739–758. <https://doi.org/10.1177/1748048510380813>
27. Maier S (2010) All the news fit to post? comparing news content on the web to newspapers, television, and radio. *J Mass Commun Quart* 87(3-4):548–562. <https://doi.org/10.1177/107769901008700307>
28. Masoodian M, Koivunen L (2018) Temporal visualization of sets and their relationships using time-sets. In: *Proceedings of the 22nd International Conference Information Visualisation, IV '18*, pp 85–90. <https://doi.org/10.1109/IV.2018.00025>
29. Rodgers P (2014) A survey of Euler diagrams. *J Vis Lang Comput* 25(3):134–155. <https://doi.org/10.1016/j.jvlc.2013.08.006>
30. Rodgers P, Stapleton G, Chapman P (2015) Visualizing sets with linear diagrams. *ACM Trans Comput-Hum Interact* 22(6):27:1–27:39. <https://doi.org/10.1145/2810012>

31. Sadana R, Major T, Dove A, Stasko J (2014) OnSet: a visualization technique for large-scale binary set data. *IEEE Trans Vis Comput Graph* 20(12):1993–2002. <https://doi.org/10.1109/TVCG.2014.2346249>
32. Sheehan S, Masoodian M, Luz S (2018) COMFRE: a visualization for comparing word frequencies in linguistic tasks. In: *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI '18*. ACM, pp 42:1–42:5. <https://doi.org/10.1145/3206505.3206547>
33. Strömbäck J, Luengo OG (2008) Polarized pluralist and democratic corporatist models: A comparison of election news coverage in Spain and Sweden. *Int Commun Gaz* 70(6):547–562. <https://doi.org/10.1177/1748048508096398>
34. Tufte ER (2000) *Visual explanations: images and quantities, evidence and narrative*, 4th printing with revisions edn. Cheshire: Graphics Press
35. Tufte ER (2001) *The visual display of quantitative information*, 2nd edn. Graphics Press, Cheshire

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Laura Koivunen-Niemi** is an information designer and alumni of the Aalto University School of Arts, Design and Architecture. She completed her M.A. in Visual Communication Design with an emphasis on Information Design in 2018. She is particularly interested in various methods of data scraping, legal aspects of data use, and design solutions centered around data integrity.



**Masood Masoodian** is a professor of Visual Communication Design at Aalto University. He has a Ph.D. from the University of Waikato, New Zealand. His research interests include visualization, interactive media, and interaction design, with particular interest in designing interactive visualizations to provide effective means of understating information by ordinary people in areas such as health and well-being, energy, and sustainability.