Aalto University
School of Arts, Design
and Architecture

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Master of Arts thesis abstract

Author Laura Matilda Koivunen

Title of thesis Narrative variance: visualizing patterns in media coverage

Department Media

Degree programme Visual Communication Design – Information Design

Year 2018 **Number of pages** 66 **Language** English

Abstract

The media plays an increasingly large role in shaping social reality, and even small shifts in its narrative content or tone can have widespread repercussions in the public's perception of past and present phenomena. Being able to track changes in media coverage over time, particularly visually, could have many conceivable applications and offer the potential for aiding social change in journalism. This case study explores how data visualization could be used to examine differences in media narrative patterns over time and across publications. The findings indicate that while there are many existing means of visualizing patterns in such narrative data on a timeline axis, few if any address the aspect of co-occurrence of variables. Comparing co-occurrence chronologically, particularly when applied to word and topic choices in media coverage, can shed more light on currents in public opinion than simply counting the occurrence of terms independently. Furthermore, the findings suggest that visualizing such patterns in this case could be best accomplished using a form of set visualization, specifically a simplified vertical version of linear diagrams repeated horizontally across parallel timeline axes. This case study also outlines the methods, ethical considerations, and examples of employing such a visualization prototype using a sample dataset of full text news articles.

Keywords data visualization, linear diagram, sets, timeline, media, news, scraping, data mining

Aalto University
School of Arts, Design
and Architecture

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Master of Arts thesis abstract

Author Laura Matilda Koivunen

Title of thesis Narratiivin variaatio: mediakertomusten visualisointi
(orig. Narrative variance: visualizing patterns in media coverage)

Department Media

Degree programme Visual Communication Design – Information Design

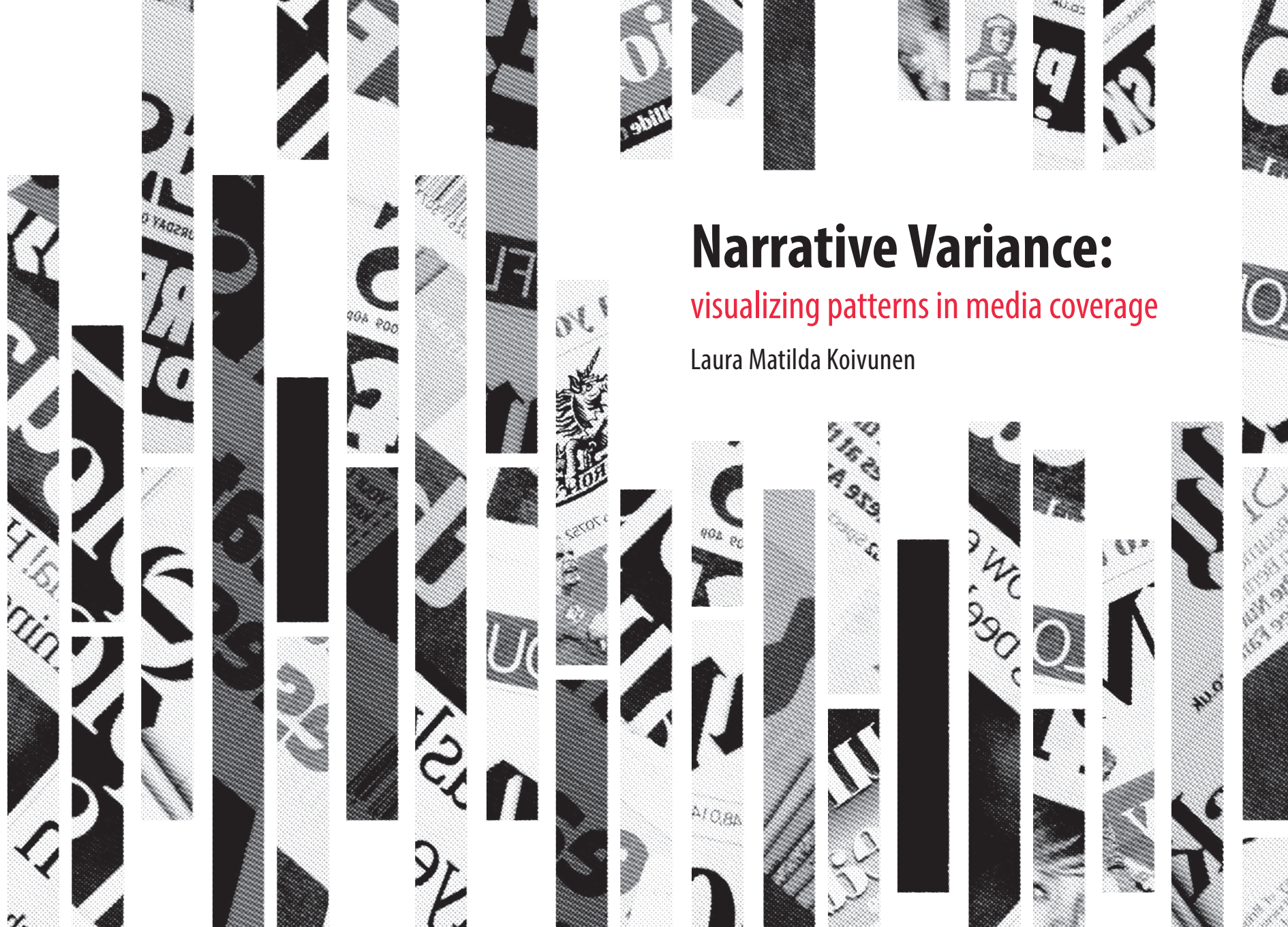
Year 2018 **Number of pages** 66 **Language** English

Abstract

Medialla on yhä suurempi rooli yhteiskunnan todellisuuden tuottamisessa, ja jopa pienet muutokset sisällössä voivat laajalti muokata yleisön käsitystä menneistä ja nykyisistä ilmiöistä. Mediasisältöjen muutosten seuranta, erityisesti visuaalisesti, soveltuisi moneen tarkoitukseen ja voisi edistää vastuullisen journalismin kehitystä ja käyttöä yhteiskunnassa. Tässä tapaustutkimuksessa selvitetään, miten tiedon visualisointia voitaisiin käyttää tutkimaan eroja mediakertomuksissa ajan myötä eri julkaisuissa. Tulokset osoittavat, että vaikka olemassa olevia keinoja vastaavan tiedon visualisointiin löytyy, yksikään ei tuo esille muuttujien samanaikaisuuden näkökulmaa. Samanaikaisuuden vertailu kronologisesti, erityisesti sana- ja aihevalintoihin sovellettuna mediasisällön osalta, voi paremmin valaista yleisen mielipiteen virtoja kuin yksittäisten sanavalintojen laskeminen. Lisäksi havainnot viittaavat siihen, että tällaisten mallien visualisointi voitaisiin parhaiten toteuttaa käyttämällä joukko-opin visualisointeja, erityisesti lineaaristen kaavioiden yksinkertaistettua vertikaalista versiota rinnakkaisilla aikajana-akseleilla. Tässä tapaustutkimuksessa esitetään myös menetelmät, eettiset näkökulmat ja esimerkit tällaisen visualisointiprototyypin tuotosta ja käytöstä uutisartikkelidataa hyödyntäen.

Keywords tiedon visualisointi, lineaarinen kaavio, joukko-oppi, aikajana, media, uutisointi, skreippaus, tiedon louhiminen


Aalto University
School of Arts, Design
and Architecture



Narrative Variance:

visualizing patterns in media coverage

Laura Matilda Koivunen

Narrative Variance: visualizing patterns in media coverage

A”

Laura Matilda Koivunen
Visual Communication Design – Information Design
Department of Media M.A. Thesis 2018
Aalto School of Arts, Design and Architecture

Abstract

The media plays an increasingly large role in shaping social reality, and even small shifts in its narrative content or tone can have widespread repercussions in the public's perception of past and present phenomena. Being able to track changes in media coverage over time, particularly visually, could have many conceivable applications and offer the potential for aiding social change in journalism. This case study explores **how data visualization could be used to examine differences in media narrative patterns over time and across publications**. The findings indicate that while there are many existing means of visualizing patterns in such narrative data on a timeline axis, few if any address the aspect of co-occurrence of variables. Comparing co-occurrence chronologically, particularly when applied to word and topic choices in media coverage, can shed more light on currents in public opinion than simply counting the occurrence of terms independently. Furthermore, the findings suggest that visualizing such patterns in this case could be best accomplished using a form of set visualization, specifically a simplified vertical version of linear diagrams repeated horizontally across parallel timeline axes. This case study also outlines the methods, ethical considerations, and examples of employing such a visualization prototype using a sample dataset of full text news articles.

Keywords: *data visualization, linear diagram, sets, timeline, media, news, scraping, data mining*

Table of Contents

FIGURE 1 – Google image searches for "immigrant" in Finnish (black) and Polish (red)

1. Introduction

Data visualization can no longer be termed an "emerging field," however, its full potential is far from 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, data visualization's potential powers of persuasion as a key element of data journalism could shed light on social phenomena that have proved divisive within society.

For instance, viewpoints within the European Union have arguably always diverged greatly on any given subject, particularly on immigration into the region in recent years. These varying viewpoints were especially tangible at a 2016 Erasmus "Living Lab" study trip to Athens, intended in an experimental think tank capacity for students from various member states to discuss issues around immigration. It became painfully apparent that there were some distinct biases and preconceptions about immigrants as people, their motives in immigrating, their impact on society, and so on. Some of these opinions were so markedly different, that any sort of constructive discussion on the subject devolved into a series of altercations. The question arose how people from a fairly constrained geographical area, with relatively similar cultural backgrounds, could hold such opposing views on an issue? Furthermore, how could a group of people with such dissimilar views ever discuss these issues without understanding the reasons behind these different viewpoints? How could the European Union ever come together and discuss these issues in a civil manner, if a group of supposedly like-minded students could not overcome this in an informal setting? There was a clear need to explain to each other the underlying cultural narrative ongoing in each student's native country, and a hypothesis emerged that these might be examined through each country's media coverage, as media debatably both perpetuates and reflects discourse in society.

As a simple experiment, a VPN-aided[1] Google Image Search was conducted to see how media in each country showed immigrants, using the te word for "immigrant"[2] in each respective language, an IP address originating in the country of choice, and the particular country's own Google portal (e.g., google.fi or google.pl). The image results were shockingly different FIGURE 1, it was clear that the narratives spun by each country's media varied drastically. On one hand immigrants were being portrayed in Poland as an encroaching threat of unwashed masses, while they were shown as happy contributing members of society in Finland. It was no surprise that citizens from different member states could not agree on this subject if their media at home perpetuated such different views, harmful stereotypes even.

This lead to another small-

1 A VPN (Virtual Private Network) was employed to ensure that search results were not tailored to the user's physical IP address location.

2 Search terms were derived from the European Migration Network's Asylum and Migration Glossary 3.0

scale exploratory project looking further into the pictures chosen to portray immigrants in news media, this time comparing news in the U.K. and Finland FIGURE 2. The findings were not particularly significant, though some lack of impartiality could be detected in news reporting by their choice of imagery. Nevertheless, this laid the groundwork for the study at hand.

From the exploratory project it became clear that image analysis on a larger scale required either an inordinate amount of time to complete manually, or a level of sophisticated machine learning as yet inaccessible to general members of the public. At the time, the most feasible means of conducting a relevant study appeared to be an analysis of the textual news content, as the technical requirements for such analysis were significantly lower. Moreover media narratives could just as well (if not better) be quantified and analyzed from word and topic choices.

From this premise, the case study set out to explore these variations in media narratives, and more specifically to answer the question of how data visualization could be used to examine differences in media narrative patterns over time and across publications. The aim in this was not to expose particular bias in media, nor to imply direct causality between media events and changes in narrative, but simply to report findings in an effective way that could be used as a further tool for media analysis and discussion on the subject.

Not only could this type of visualization set the table for discussing different media viewpoints for various purposes, it could also be useful in a media oversight capacity. Being able to visually track media trends in a particular publication, or on a particular subject, could be an invaluable tool for accountability and oversight, especially for citizens in countries with unregu-

FIGURE 2 – Excerpt from an initial project showing immigrant-related image use in media, topics such as violence are highlighted in color.

lated press where news may be sensationalized more often than not. Even in a country like Finland where media is held accountable for producing impartial news through the Council for Mass Media (Finnish Julkisen Sanan Neuvosto) and the Finnish Communications Regulatory Authority, Ficora,[3] being able to determine whether a national narrative is drifting in a direction not "in accordance with journalistic principles" as is required by the national Guidelines for Journalists, could be an important aspect of media oversight. All this has immense analytical potential, but limited by the lack of big data-driven tools at the moment.[4]

Data visualization in this sense has a unique opportunity to contribute to discourse about media use, and furthermore to the media research field 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 exponentially faster with an appropriate visualization tool. Traditional methods of communicating media analysis information arguably fall short in their ability to simplify and communicate patterns such as this in large datasets.[5] The key specifically lies in the ability to automate the processing and simplification of large datasets into something visual and comprehensible. This task could of course be approached from many angles.

In this study we find that although there are countless ways to visualize simple patterns in media narratives, one method which brings new insights into, not only the occurrence of variables in media, but their co-occurrence in a large dataset, utilizes vertical linear diagrams across parallel horizontal timelines. The resulting visualization type is debatably something new in the field, as no other visualization types were identified at this time that addressed this particular angle of analysis. This work does not address the quantifiable effectiveness of this new visualization compared to other variations —in part because there is little if anything to compare to— but encompasses a case study that explores the creation of this prototype implementing a particular set of methods, as well as ethical implications and use examples.

3 Jyrkiäinen, Jyrki. "Media landscapes: Finland." European Journalism Centre.

4 Julkisen Sanan Neuvosto. "Journalistin Ohjeet Ja Liite."

5 Jensen, Klaus Bruhn. *A Handbook of Media and Communication Research*, 262-263.

2. Scope

Before selecting a dataset to visualize changing media narratives, it was necessary to consider what sort of granular data could show currents in narratives. Narratives arguably stem from words and the sentiments that could be inferred from them; so gathering the bulk of the "words" used by media, that is to say the contents of the news, seemed most pertinent for this study. The sections below outline the selection of this dataset, which consisted of articles from three U.K. newspapers over the years of 2010–2016. In the interest of conducting a media study in the most scientific way possible, and because data (and its integrity) should always come first in a work of data visualization, considerable effort was afforded to the documentation of the data itself. It is imperative to especially document the data scope selection, as the resulting dataset alone —not to mention the visualization of such data— can cause the viewer to draw the wrong conclusions if the selection involved any amount of "cherry picking." Particularly in the age of fake news and when discussing a contentious issue such as immigration, it is necessary to provide the greatest possible transparency in the practice of (what we might in this case call) data journalism. Furthermore, the motive in explaining this process is not only to demonstrate due diligence in data gathering methods, but also to provide some guidelines should it be necessary to replicate this visualization in some form.

2.1 Country and language

For practical purposes of small-scale language processing and a manageable data scope, the news chosen for analysis was limited to sources based in the United Kingdom. It would nevertheless have been especially interesting to compare how narratives varied between media in different countries on a given subject, as was the initial premise of this study. The work in the end necessitated a single language source, and English language news could be most easily acquired, analyzed and understood by both the persons involved in the development of the visualization as well as on the observing end of the spectrum. Furthermore, without significant linguistic training or sophisticated language processing tools, adding other languages into the mix would only bring the integrity of the dataset and analysis into question. It was also hypothesized that by using the U.K. instead of another English-speaking country such as the U.S. it would be possible to follow European media narratives such as the migration crisis from a more insular vantage point.

2.2 News media

Traditional newspapers in their various forms, though declining in circulation, still remain a major source of information in today's society. Increasingly in an effort to remain relevant, news companies have been migrating to a more active web presence, publishing news both on their own sites as well as social media extensions. It was postulated that with more

than half of Britons getting their news online already in 2013, surveying the online content of daily newspapers could be a feasible means of tracking media narrative currents in the 2010s.[6] A majority of people reported in a British National Readership Survey that they read the online versions of newspapers much more often than any print version, with the exception of a select few titles such as the *Sunday Times* and *Metro*.[7] Television, of course, remains particularly prevalent as a news source in Britain according to a 2013 OfCom report,[8] but for the purposes of this study it was deemed that word analysis would be more feasible from a written record rather than audiovisual material.Televised news was nevertheless surpassed by its online competitors in 2016 as the most used news source, with over 70% of participants reporting online news consumption (including social media-based) within the survey period.[9]

6 Sweney, Mark. "More than half of Britons access news online." *The Guardian*.

7 OfCom. "News consumption in the UK."

8 Ibid.

9 Nielsen, Rasmus Kleis. "Where do people get their news? The British media landscape in 5 charts." Medium.

Top circulating print publications in the U.K. 2010–2016

	2016	2015	2014	2013	2012	2011	2010
○ Sun	1,787,096	1,978,702	2,213,659	2,409,811	2,582,301	3,001,822	3,006,565
○ Metro	1,348,033	–	1,362,893	–	–	–	–
● Daily Mail	1,589,471	1,688,727	1,780,565	1,863,151	1,945,496	2,136,568	2,120,347
Evening Standard	898,407	877,532	805,309	695,645	699,368	704,008	601,960
Daily Mirror	809,147	922,235	992,256	1,058,488	1,102,810	1,194,097	1,218,425
○ Times	404,155	396,621	384,304	399,339	397,549	457,250	508,250
Daily Star	470,369	425,246	489,067	535,957	617,082	734,311	779,376
● Daily Telegraph	472,033	494,675	544,546	555,817	578,774	651,184	691,128
Daily Express	408,700	457,914	500,473	529,648	577,543	639,875	674,640
i	271,859	280,351	298,266	293,946	264,432	133,472	N/A
Financial Times	198,237	219,444	234,193	275,375	316,493	383,067	390,315
● Guardian	164,163	185,429	207,958	204,440	215,988	279,308	302,285
...	...	...	...	...	...	...	...

2016 TOP 3 ● Tabloid publications ● Broadsheet publications ○ No online archive

TABLE 1 – Top-circulating print publications in the U.K. 2010–2016, average circulations for January of each year, marked with publications chosen for the study.

The particular publications featured in this dataset were initially identified as some of the top circulating newsprint publications of 2016 TABLE 1.[10] As a means of ensuring a variety in both objective and more sensationalist news coverage it was intended that the dataset should include the three top circulating broadsheets[11] as well as the three top circulating tabloids.[12] It was necessary however to employ a form of convenience-sampling and limit the selection to three publications, as several of the top circulating newspapers had no freely accessible online archive. Fortunately however, the three remaining titles enjoyed the most online readers out of the original six selections, as outlined in TABLE 2.[13] As online news was the primary focus of this study, the online readership scores understandably outweighed any physical print circulation statistics in sample selection.

10 Wikipedia. "List of newspapers in the United Kingdom by circulation."

11 "Broadsheet: a newspaper with a large format, regarded as more serious and less sensationalist than tabloids." *Oxford Dictionaries* s.v. "Broadsheet."

12 "Tabloid: a newspaper having pages half the size of those of the average broadsheet, typically popular in style and dominated by sensational stories." *Oxford Dictionaries* s.v. "Tabloid."

13 National Readership Survey. "Newsbrands: Print/PC."

Top [online] circulating publications Oct. 2015 – Sep. 2016

Publication	Print	PC (online)	Print + PC
Daily Mail/The Mail on Sunday	12,292,000	7,418,000	17,494,000
Daily Mail	10,068,000	7,418,000	15,604,000
The Guardian/ The Observer	4,594,000	6,972,000	9,965,000
The Guardian	4,037,000	6,972,000	9,593,000
The Daily Telegraph/ The Sunday Telegraph	4,396,000	6,740,000	10,047,000
The Daily Telegraph	3,819,000	6,740,000	9,569,000
Daily Mirror/ Sunday Mirror	6,802,000	4,374,000	10,548,000
Daily Mirror	5,902,000	4,374,000	9,719,000
Sunday People	727,000	4,374,000	5,039,000
. . .	. . .	. . .	. . .

TABLE 2 – Top online circulating publications in the U.K. October 2015–September 2016, publications chosen for the study are highlighted.

2.3 Timespan

The timespan of 2010-2016 was chosen particularly with topics such as immigration and Brexit[14] in mind. It was initially hypothesized that changes in narrative might be particularly apparent on certain migration- / immigration-related topics over those years due to the inflow of asylum-seekers and migrants into the European Union during that time. It was additionally hypothesized that the migration narratives might possibly culminate in a marked peak at the 2016 Brexit vote.

While the data scope was set and visualization built around these (albeit quite broad) pre-defined topics, it should be noted that the dataset in itself does not exclude the possibility of examining the material for narrative patterns on other subjects as well. Articles were indiscriminately included in the dataset irrespective of whether or not they mentioned these particular topics.

2.4 Categories

Including *all* articles from the specified time period proved to be untenable however, due to the sheer volume of content disseminated by the three chosen publications — particularly by *The Daily Mail*, which boasts thousands of articles per day. Subsequently article selection was streamlined to include only topics deemed by the publication itself to be some form of news or politics, determined from an extensive list of source-specific categories outlined below.

14 "Brexit: The withdrawal of the United Kingdom from the European Union." *Oxford Dictionaries* s.v. "Brexit."

The categories chosen from each publication for the dataset were as follows:

The Daily Mail

Included: *news, wires*
Excluded: *showbiz, sports, "femail," Australia, U.S., India, fashion, travel, video, health, money*

The Telegraph

Included: *news, politics*
Excluded: *video, sport, culture, travel, lifestyle, women, men, fashion, luxury, tech, film, expat, sponsored, promotions, food & drink, motoring, education, gardening, finance, money, business*

The Guardian

Included: *UK news, world news, politics*
Excluded: *life and style, society, education, law, Scotland, Wales, media, Northern Ireland, travel, music*

2.5 Data fields

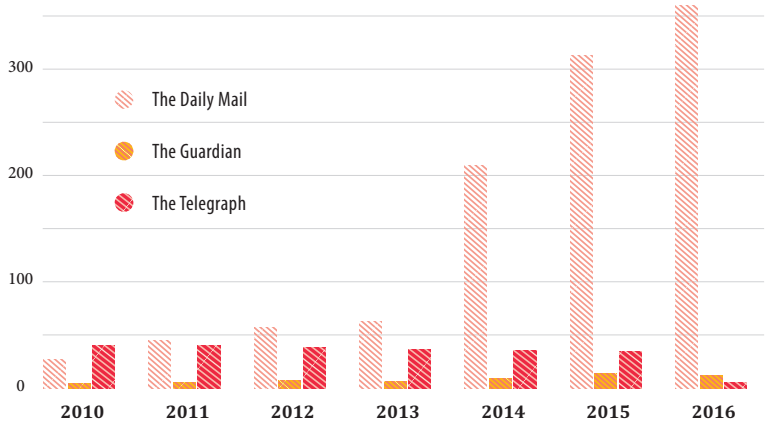
The data fields selected for collection were initially as extensive as possible, and included the publication name, article title, publication date, URL, full text, keywords, main image URL, full page HTML, and a unique database ID. Much of this information proved to be superfluous for the basic needs of the visualization however; in the end only the publication

Sample article data

id	5a7052f4257e060c6ffc3183	**Date**	2016-12-30
Title	Police search for man who damaged Skye dinosaur footprints		
Image	https://i.guim.co.uk/img/media/b1161e77c23504ee7a8a8fa008b5e105c2600d82/0_278_4106_2464/master/4106.jpg		
URL	https://www.theguardian.com/uk-news/2016/dec/30/police-search-for-man-who-damaged-skye-dinosaur-footprints		
Text	Scotland Police search for man who damaged Skye dinosaur footprints Man sought after he poured plaster into the fossilised footprints, which are thought to 165m years old, at Staffin beach The footprints are a tourist attraction on the island. Photograph: Murdo MacLeod for the Guardian . . .		
Keywords	Skye, dinosaur, fossils, police, search . . .		
HTML	<!DOCTYPE html><html id="js-context" class="js-off is-not-modern id--signed-out" lang="en" data-page-path="/uk-news/2016/dec/30/police-search-for-man-who-damaged-skye-dinosaur-footprints">...		

TABLE 3 – *(above)* Sample of data gathered per article

FIGURE 3 – *(right)* An outline of the dataset used in this study



name, publication date, and full text element were used from each article. Some of the other elements could be utilized in future extensions or adaptations to provide granular article-level detail though. Below in TABLE 3 is an example of one article entry in the dataset.

The full extent of the acquired data was in the end quite surprising. For three publications over the years of 2010–2016, a grand total of 1,344,806 articles were collected. The breakdown of articles per source and year is outlined in FIGURE 3, and is in itself quite telling about publication trends and behaviors. As can be seen, the dataset does unfortunately skew heavily in the direction of *The Daily Mail*, due to the sheer volume of their articles. This was hopefully compensated for in the visualization by addressing monthly data separately by publication instead of being averaged together. It should also be noted that the annual article count for *The Telegraph* is so low in 2016 due to the unfortunate discontinuation of their online archive in the spring of that year.

3. Materials and methods

While the data and its selection are paramount in this study, the methods of its implementation are afforded a somewhat secondary status in documentation. An entire thesis could be written on the technical aspects alone, but only a cursory discussion of the working pipeline will be covered in this section, as this is primarily a treatise on visualization. Nevertheless full code documentation is available on a GitHub page designated for that purpose,[15] and sample pages of code are provided in the Appendix.

It should be mentioned that there are many ways these same goals could be accomplished (see for example the pipeline in Pierre Bellon's "Islam, media subject")[16] and the methods proposed and employed here are by no means the only options. As much of the implementation was a personal learning process building on code samples generously afforded by the Antwerp-based Experimental Media Research Group (EMRG),[17] the means here are adequate in meeting the ends but likely do not reach their full intended potential.

3.1 Scraping

The primary method of data acquisition consists of automating scripts to crawl a series of pages online, then copy specific content therein to a centralized database; this process is hereafter referred to as "scraping." The ethics of these methods will be further discussed in the following *Legality and Ethics* chapter.

The primary script template used in scraping the dataset is written in Python, and utilizes the BeautifulSoup[18] and Newspaper3K[19] python libraries to identify and extract articles in conjunction with the pymongo[20] distribution to pass the scraped data to the MongoDB[21] database. The scraping script worked in two consecutive loops, first using BeautifulSoup to systematically gather a list of URLs to be scraped from the archive pages, and then on a second pass visiting each URL and using Newspaper3K to copy the variables identified as article elements, see Figure 4. The template script was varied for each publication in turn to account for differences in the various archive pages, as well as included/excluded news topics (see previous chapter for details) and other discrete nuances.

The scripts scraped only the articles from the archive pages that were outlined within the scope category, for example articles about "news," not about "showbiz." The websites' URLs were employed to this end to identify an individual article's category, as fortunately each publication had chosen to construct the URL using category tags. A URL containing

15 GitHub code documentation https://github.com/lauramatilda/newsScraping

16 Bellon, Pierre. "Islam, media subject: How to quantify the perception of Islam in the media." Data Driven Journalism.

17 Experimental Media Research Group (EMRG) http://emrg.be/

18 BeautifulSoup https://www.crummy.com/software/BeautifulSoup/

19 Newspaper3K http://newspaper.readthedocs.io/

20 Pymongo https://api.mongodb.com/python/current/

21 MongoDB https://www.mongodb.com/

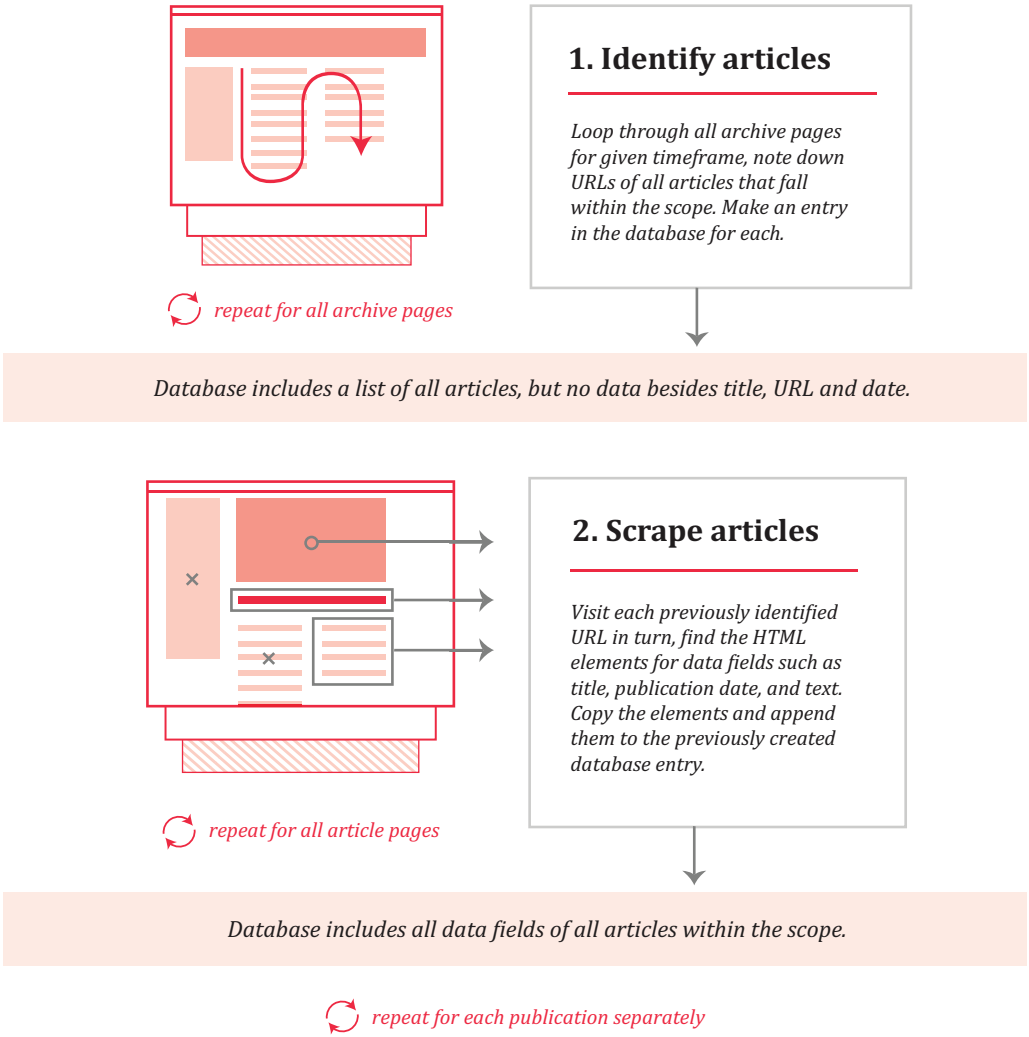


FIGURE 4 – The article scraping process.

the substring `"theguardian.com/uk-news/..."` could for instance be identified as the category "UK News," and so on.

Furthermore the scripts were given a specific date range from which to scrape, also fortunately queriable through the archive page URL for the most part. *The Daily Mail* archive page URLs for example follow a simple pattern of `".../sitemaparchive/day_%s.html"` where `%s` is a string format of a date. The first part of the script would loop over archive pages until it reached the end of the scope (e.g., January 1, 2017) and then the second part of the script would loop through the list of articles until each one was fetched into the database.

3.2 Database

A MongoDB NoSQL[22] database was used in this study to house the scraped articles. A NoSQL database proved to be particularly useful for this study, as the initial scope of the dataset was not known at the outset, and NoSQL is

22 NoSQL (no Structured Query Language) https://en.wikipedia.org/wiki/NoSQL

particularly scalable unlike its SQL[23] colleague. The database was run both on a local computer as well as on a private remote-hosted server, depending on the particular need. For example, initial scraping was attempted using the hosted database, but due to constant server crashes and slow response times, it was deemed necessary to scrape first to the local machine and then deploy the database in its entirety to the server. Later on during NodeBox implementation (discussed below) it was useful to refer to the hosted instance, which would also be the case in deploying a live interactive version of the visualization. Generating the vector images destined for print, on the other hand, was much faster when pulling from the local database instance. Much of this need for disparate databases stemmed of course from the volume of data, lack of data pre-processing, and slow server response times. Ideally all tasks could be executed using one hosted instance of the database.

23 SQL (Structured Query Language) https://en.wikipedia.org/wiki/SQL

3.3. API

A Flask[24] Python microframework application was deployed on a free Heroku[25] instance to serve up the Application Programming Interface (API)[26] used in this study. In short, the Flask application pulled articles from the MongoDB database (either from the local or server instance) and returned the queried data in a plain text format that could be used in the next step of the visualization. By default the API query returned all available data per article, but tailoring the query to only select smaller subsets of article data fields opened up different possibilities for visualization. Some API variants considered for use in the study included article keywords, articles that only dealt with certain topics, articles from a specific timeframe, full text for a limited number of articles, and so on. For this particular application it was deemed necessary to output one publication's full text articles one month at a time, with no topic restrictions, and excluding all other data fields such as titles, images, or HTML. One limitation was unfortunately necessary to implement: restricting the number of articles output per month. Ideally all articles per publication per month could be included, but again a combination of slow response times and an overwhelming amount of unprocessed data being piped into the visualization tool necessitated limiting the number of articles used to at most 1000 at a time. Response times were further improved by limiting the articles to 750, 500, and finally even to 250. To preserve some semblance of a representative sample however, it would be appropriate to include as many articles as possible if the full set is not attainable.

24 Flask http://flask.pocoo.org/

25 Heroku https://www.heroku.com/

26 API (Application Programming Interface) https://en.wikipedia.org/wiki/Application_programming_interface

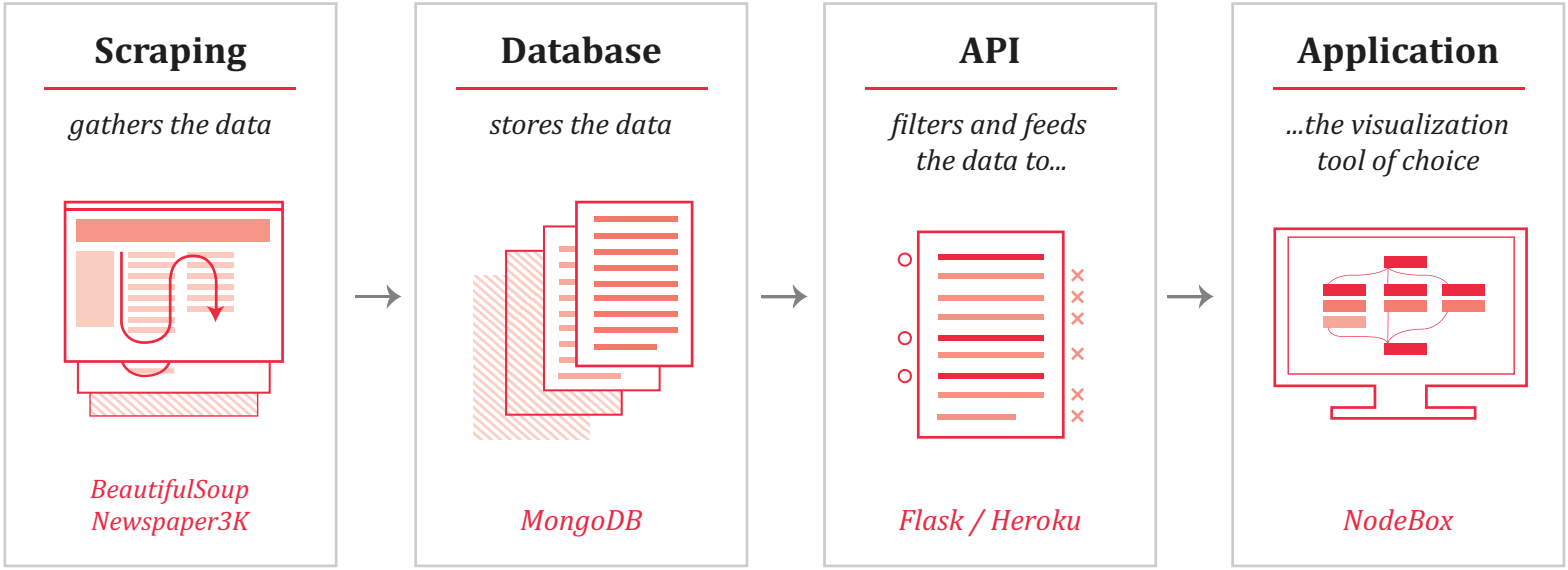


FIGURE 5 – The production pipeline.

3.4 Application

The final step of the visualization was integrating the API into a visualization tool of choice to produce the prototype. This particular prototype was built using the EMRG's NodeBox Live[27] tool, a node-based visualization platform that can be run in a browser. It should be noted that the prototype could have been implemented in a variety of ways, for example coding it using the D3 javascript library,[28] however NodeBox Live proved to be a versatile and intuitive tool with relatively easy API integration. One particular strength of the tool proved to be the ease at which design elements could be exported out in vector format for use in Illustrator and InDesign, as well as the ability to experiment with an interactive version of the same prototype. The actual specifics of the visualization implemented with NodeBox will be discussed in the forthcoming *Visualization Development* chapter.

27 NodeBox Live https://nodebox.live/

28 D3 https://d3js.org/

4. Legality and Ethics

Ethical considerations are often a source of great contention in journalism and the media. Coincidentally, some of the most notable ethical considerations affecting designers of data visualizations are the same that affect traditional journalists. Communicating evidence in one form or another to an audience, whether it is in a textual summary or a visual form, involves a series of ethical decisions,[29] beginning with what to include or exclude, how to summarize and report the data without compromising its core facts with personal bias, transparent documentation of methods, and so on. The number of ways a journalist or designer can intentionally or unintentionally distort data (and thereby jeopardize the principals of the entire endeavor) are staggering, and many treatises on this subject have been written since the beginning of journalism. Nevertheless, a particular set of ethical and legal considerations which are not well documented in the literature as yet, and are particularly applicable to this case study, are those concerning the automated gathering and use of data. Extensive discussion on this topic may seem pedantic in a visualization case study, but is highly relevant for designers who may employ these data acquisition methods.

As mentioned in the previous *Scope* and *Methods* chapters, for the purposes of developing this visualization, a substantial dataset of articles was harvested. The articles were copied from their original locations —as opposed to referring to them directly *in situ*— in order to obtain a sample with consistent field sets and formatting across the entire scope. This local database also enabled the implementation of a customized API and integration with the NodeBox application. This would not have been possible with the articles in their original disparate locations. Under ideal circumstances, if the publications provided standardized APIs in the manner necessary for our purposes, the visualization could have been constructed using open data. Some third-party news aggregate APIs do exist and could be used for similar visualizations, but their temporal scope is generally limited to a single day, with article quantities ranging between 10 and 100 popular titles.[30] Unfortunately, these third-party APIs were not deemed to constitute a thoroughly representative sample, and that scraping in bulk was necessary for obtaining an adequate sample.

Gathering data —in bulk or otherwise— for journalistic purposes in itself has a particular set of rules, but as this study is conducted in an academic setting, we will examine the legal aspects of scraping for research, not for journalism (though the data might conceivably be used for both purposes). Many might argue that this bulk scraping approach to data gathering is acceptable in academia, and the Finnish Copyright Society (Kopiosto)[31] does indeed provide licensing for even

29 Tufte, Edward R. *Beautiful Evidence*, 141.

30 NewsAPI https://newsapi.org/

31 Kopiosto. "Copying in schools and educational institutions: Scope of permissible copying in teaching." Kopiosto Copyright Society.

large-scale data mining in the name of scientific research, even newspaper articles as used in this study. However, following a closer reading of the Kopiosto licensing terms, combined with a survey of European legal case history (discussed in more detail below), it seems the licensing may not cover the particular dataset used in this study.

It should be stressed at this point that the effects of gathering and temporarily retaining this dataset for research purposes should not in our estimation unduly provoke legal action by constituting a hindrance, or causing financial hardship — and a more lenient legal eye might even find these methods permissible. It is nevertheless important to be mindful of this aspect of research and examine in detail the legal and ethical implications of the chosen methodology. To this end, it is also pertinent to survey some of the different contexts (i.e., legal jurisdictions) in which these methods might be addressed and challenged, as well as the various points of contention and possible legal exceptions.

4.1 U.S. and European court decisions

The legality of scraping data from news websites, or for that matter any websites in large quantities, is in many respects considered tenuous. After all, the companies in question have used their own budgets and resources to create the content now being harvested for someone else's purposes, and it is in many cases their sole product and source of income aside from advertising revenues. Some companies have in fact argued in court that scraping their publicly available data[32] is an act of "hacking" and thereby intellectual property theft. In the U.S. the Computer Fraud and Abuse Act (CFAA), which outlaws "access [to] a computer without authorization or [in excess of] authorized access," has been cited as the legal backing for restricting scraping access.[33] One such example is found in the U.S. case *hiQ Labs, Inc v LinkedIn Corporation,* in which the startup firm HiQ Labs challenged LinkedIn's rights to block their data scraping efforts. The judge sided with the startup, ruling that the information was publicly available, and therefore LinkedIn had no right to stop another company or individual from accessing that data.[34] The decision referenced legal scholar Orin Kerr, who stated that allowing website owners to arbitrarily deny certain access could lead to discriminatory practices such as blocking based on political affiliations, race, or gender.[35] Kerr postulates in his 2015 landmark paper "Norms of Computer Trespass" that anything published without password protection on the internet has already been made public, and that accessing any such content cannot exceed "authorized access" and thereby violate the CFAA, since the website provider has already authorized everyone and anyone to access that content.[36] Based on these judgments, the implications for accessing and scraping publicly

32 In this context "publicly available data" refers to any data accessible to the general public on the internet, not "open data," which refers to data released for public use with specific licensing terms with the intent of its reuse.

33 Lee, Timothy B. "Court rejects LinkedIn claim that unauthorized scraping is hacking." Ars Technica.

34 United States District Court: Northern District of California. *hiQ Labs, Inc v LinkedIn Corporation*.

35 Lee, "Court rejects LinkedIn claim…"

36 Kerr, Orin S. "Norms of Computer Trespass." SSRN.

available data, at least within the U.S., seem to be favorable. Alas, this is not the case everywhere.

In the European sphere, the legal status of scraping is somewhat more complicated. The European Commission moved in 2016 to expand copyright law to allow for an "exception that would permit researchers to analyze on a large scale scientific data to which they have lawful access."[37] The proposal would specifically make allowances for Text and Data Mining (TDM) by public and private universities and research institutions with a public interest goal. From personal anecdotal evidence, it seems at least some scholars have already taken this as an implicit mandate for scraping, but the decision itself is still pending in the European Parliament. Aside from this, the European courts themselves have not been as sympathetic to data scraping efforts. They sanctioned the right to gather data to a certain extent in their 2014 decision *Ryanair Ltd v PR Aviation BV,* but provided a "loophole" for site owners to forbid future scraping.[38] The court ruled that while scraping of a database falling outside the 1996 Database Directive (i.e., databases of public data or non-original works) could not be forbidden outright, a company or site might restrict scraping by outlining the allowed uses of their database in the site's Terms of Use.[39] As a result, one might imagine that many companies updated their Terms of Use to specifically forbid any scraping of their public database. For example, *The Guardian* site Terms of Service allow for "personal and non-commercial use" of their content, but also unfortunately state that "except as expressly authorized by *The Guardian*, you are not allowed to create a database in electronic or paper form comprising all or part of the material appearing on *The Guardian* Site."[40] This is particularly unfortunate because the methods of creating the visualization do, as an intermediary step within the process, involve the creation of a database of the scraped articles. This is also the reason why the Finnish Kopiosto licensing for university staff and students debatably cannot be used to defend these scraping efforts, as those licenses allow for copying material for the purposes of research only if no separate license or usage terms agreement exists for the material.[41] *The Guardian*'s Terms of Use, arguably, are just such a separate license, and possibly supersede the Kopiosto licencing.

In this sense, for the purposes of furthering media transparency and data visualization's role in facilitating discussion on the subject, this case study is using data scraping as an academic form of civil disobedience. All this is pending a definitive decision from the European Commission about bending copyright law in favor of scraping for research; in the meantime the legal implications of scraping for these purposes in Europe remain unclear.

37 European Commission. "Commission proposes copyright exception for researchers." European Commission: Policies, Information and Services.

38 McLean, Susan and Mercedes Samavi. "Data for the taking: using website terms and conditions to combat web scraping." Lexology.

39 Court of Justice of the European Union (CJEU). *Ryanair Ltd v PR Aviation BV*.

40 *The Guardian*. "Terms of Service."

41 Kopiosto. "Copying in schools and educational institutions: What the license does not permit." Kopiosto Copyright Society.

4.2 Use of Resources

Another wrong that might be leveled at these scraping efforts is the amount of the web host's bandwidth and other resources that have been used up in the process. Accessing a site repeatedly, sometimes tens of thousands of times per day, can cause undue stress on a server and use up bandwidth that might otherwise be allocated to regular browsing of the site. In cases such as *eBay, Inc. v. Bidder's Edge, Inc.*, a U.S. court sided with the plaintiff and ordered the scraper to pay damages for the amount of server resources used up in scraping efforts, additional man hours needed for upkeep, damages to profits caused by slow server response times, etc.[42] The nature of the accessed data in these cases does not usually Figure prominently in the grievances (though it is of course an underlying point of contention), but relies more on the long-standing U.S. legal concept of *trespass to chattels*. This ancient legal concept mandates compensation for damages caused by the use of another's personal property in such a way as to cause the owner undue harm in some way. Hypothetically in this case study, the personal property at stake would again be the publications' servers, and the harm caused by the scraping could for instance be diminished server performance leading to a need to purchase more computing power. In this case the pace of scraping was hopefully limited enough not to noticeably increase demand on any server. Such unintentional resource use, legally permissible or not, should be examined from a moral standpoint whenever undertaking scraping efforts in bulk.

42 Internet Law Treatise. "Trespass to Chattels."

4.3 Nature of data, nature of use

The question also remains whether the nature of the data makes a difference in the ethical or legal quality of its scraping — for example intellectual property vs. a user's personal data. In the aforementioned U.S. LinkedIn scraping case, the business-networking site claimed to be defending its customers' right to what information they made public, referring to hiQ Lab's projections about when an employee was likely to leave their job based on scraped LinkedIn data.[43] The data the projections were based on was indeed freely made public by each individual user, but likely the users did not think the information could be used to determine their personal employment patterns. Nor would they want their current employer to find out that they were intending to leave a particular job (though it is pertinent to consider whether an employer might not have made these same conclusions on their own, based on the employee's increased LinkedIn usage). The sensi-

43 Lee, "Court rejects LinkedIn claim…"

tive nature of the data nevertheless might have made that particular use ethically unsound, if not necessarily legally.

Aside from the nature of the data, the nature of the use could be considered a defense in some jurisdictions. For example, from a U.S. legal standpoint, since this particular case study's data is publicly available (i.e. not password-protected), and the scraping is for personal, academic, and non-commercial use only, one could assume that the scraping practice falls under the legal concept of "fair use." This concept makes allowances for copying of intellectual property for academic and non-profit purposes, compilations made from a collection of other works, etc. It should be noted however, that this notion only exists in U.S. courts, and is only applicable as an active defense in response to charges of copyright infringement; fair use cannot by any means be used as a deterrent to legal action.[44-45]

44 Smith, Kevin, M.L.S., J.D. Lecture notes from "Copyright for Educators & Librarians." Duke University, Coursera.

45 Internet Law Treatise. "Fair use."

Regardless of the jurisdiction, data origin, intentions for data use or particular methods used, it is evident from even a cursory review of legal cases that the grounds for scraping are still in a formative period, and its justifications untenable in many cases. As in many other aspects of modern society, legislation seems to lag behind technological advancements and legal precedent must be forged through the arduous process of court cases. In the absence of more consistent rulings on the subject and a consensus on a restructured copyright policy —at least within the European jurisdiction — we can only take note of the possible legal implications of research using data mining, and proceed with caution.

5. Visualization development

5.1 Review of media visualizations

Given the dataset at hand, it is pertinent to consider what aspects could be visualized to achieve the required ends without skewing the dataset. Surveying other visualizations trying to realize the same goals could conceivably yield some insights, particularly ones that visualize some aspects of news media on a temporal scale. While it is difficult, if not entirely impossible, to identify all the relevant visualizations with similar features, a survey of popular media visualizations returned the following specimens for analysis.

Traditional media study visualizations

Table 4 [46] shows some of the basic forms of visualization usually used in conjunction with formal media studies. The data presented therein represents the rate of occurrence of certain words or topics in surveyed media, suggesting alternately a trend of sensationalizing news on one hand, and marginalizing gender groups on the other. The data speaks admirably for itself and does not need excess visualization or interpretation for the viewer to gain insights into the study. The use of plain tables in these sorts of situations is sufficient, and attempts to contrive anything more complicated just for the sake of having an eye-catching chart would result in—as Tufte would put it—"chartjunk".[47]

46 Jensen, Klaus Bruhn. *A Handbook of Media and Communication Research*, 109-110.

47 Tufte, *Beautiful Evidence*, 141.

Use of names for the flu in Swedish Twitter

	H1N1	New flu	Swine flu	Others	No name
Official crisis authority	38%	18%	8%	20%	16%
Public service broadcaster	–	6%	90%	4%	–

Content analysis of review of the year 1998 BBC1

	Men	%	Women	%
British	39	56%	15	65%
American	8	11%	4	17%
European	8	11%	2	9%
Rest of the world	14	20%	1	4%
Unidentifiable	1	1%	1	4%
Politics	38	54%	7	30%
Showbiz	22	31%	12	52%
Sports	16	23%	4	17%
Enterntainment	6	9%	8	35%
Science	2	3%	0	0%
Everyday life	8	11%	4	17%
GENDER	70	75%	23	25%

TABLE 4 – Two typical media study content analysis tables

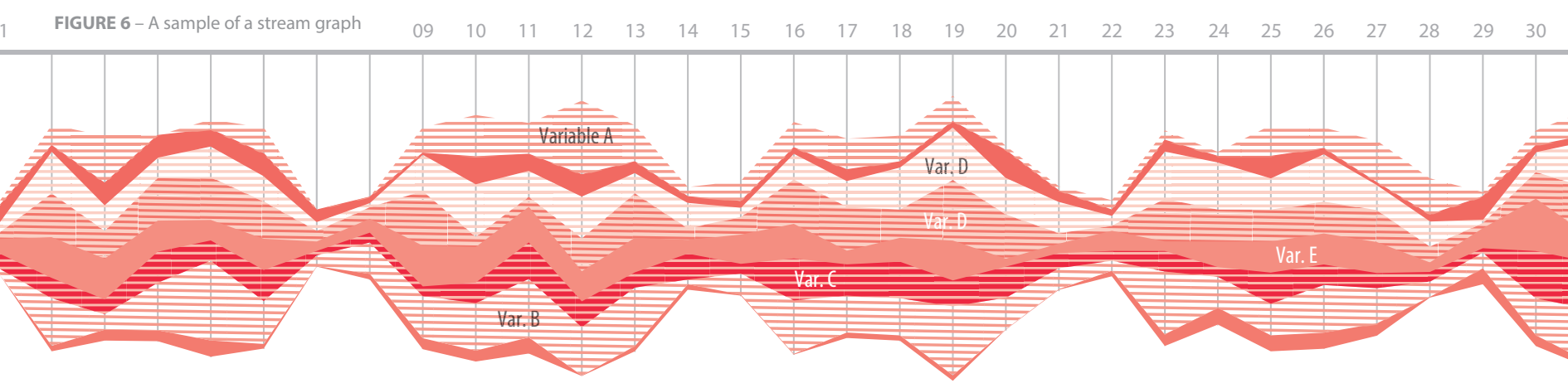


FIGURE 6 – A sample of a stream graph

Nevertheless, problems with table displays of data can foreseeably arise when the dataset is expanded to show different points in time or other additional variables, rendering the table considerably larger and more complicated. Identifying patterns over time, especially patterns in variable use correlation would prove to be considerably more cumbersome and would require a different form of visualization using a timeline element.

News coverage stream graph

One such example of a timeline element in use is found in another master's thesis on the subject of news archive visualization from 2007, which proposes many means of visualizing BBC news coverage by geographic area and keyword over time.[48] The one visualization type from Li Xin's thesis "Mapping the recent past: Visualization of online news archives" which might be most appropriate to review in this situation is the stream graph FIGURE 6 showing news coverage of a geographic area over time. The variations over time and the prevalence of certain areas in the news are particularly apparent, but stream graphs are problematic for anything more than a rough glimpse of relative data patterns. Accurate measurements of any particular geographic area, or comparisons between one point in time to another are nearly impossible without a common baseline.[49] A basic line graph over time might have served this type of visualization better.

Google Books Ngram Viewer

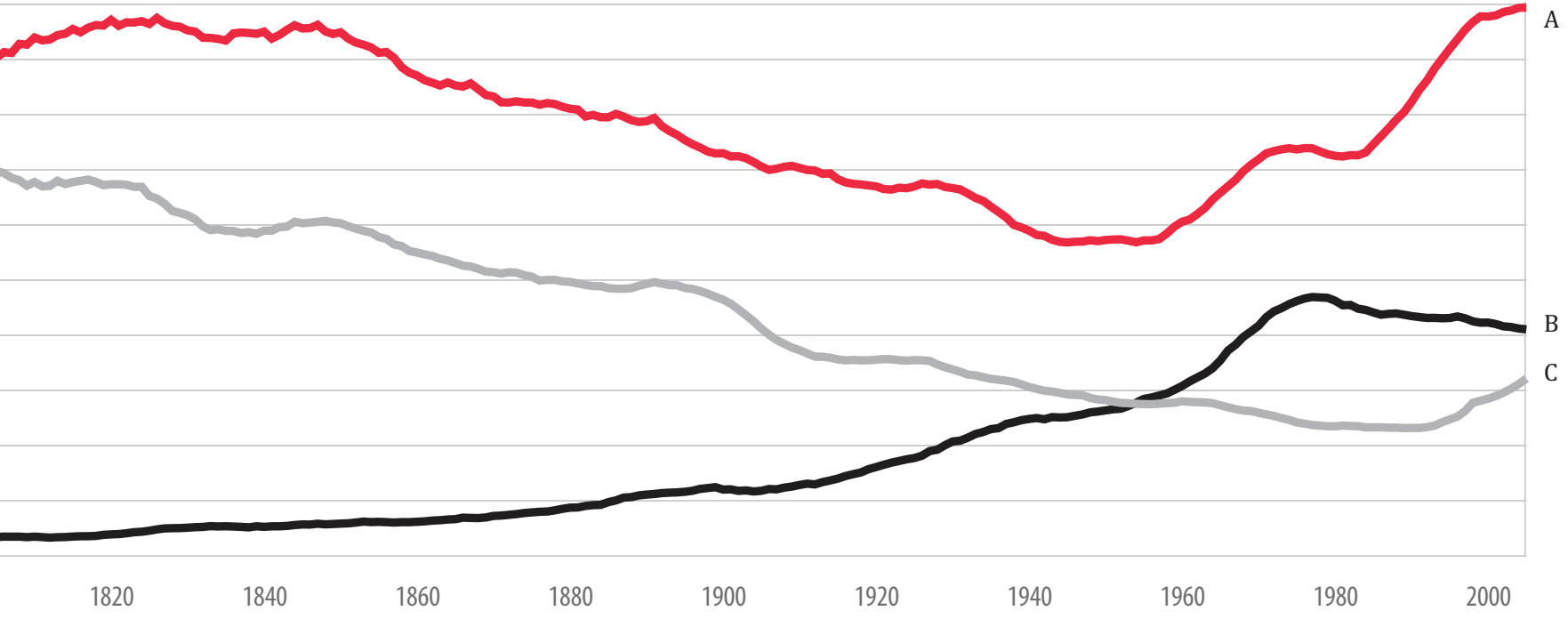
Google Books[50] uses such line graphs to show word use chronologically. The interactive user interface allows

48 Xin, Li. "Mapping the recent past: Visualization of online news archives."

49 Koponen, Juuso, Jonatan Hildén, Tapio vapaasalo. *Tieto Näkyväksi*, 191.

50 Google Books Ngram Viewer https://books.google.com/ngrams

FIGURE 7 – Line graphs, as used in Google Books Ngram Viewer



for comparison of words across time, superimposed on a horizontal timeline FIGURE 7. The setting is usually best suited for showing the continuity of data through time, though in this case a sequential bar graph showing each year's discrete measurements would maybe have represented this particular dataset more accurately.[51] There is furthermore the problem of multiple scales being combined in one visual; variations in a smaller scale line diagram can be misrepresented or obscured by those of another magnitude. Moreover, displaying the line graphs superimposed on each other seems to imply a relationship between the different words simply based on existence on the same timeline. It is easy to draw a faulty parallel: the viewer may well think that if words A and B both increase in use at the same time, then there must be a significant connection and possibly a common catalyst for change.[52] Showing line graphs together when there is no particular connection between the sources is problematic to say the least; this visualization is perhaps best used to show the changes in the use of a single word over time.

51 Koponen, Hildén, Vapaasalo, *Tieto Näkyväksi*, 190.

52 Tufte, Edward R. *Visual Explanations*, 102.

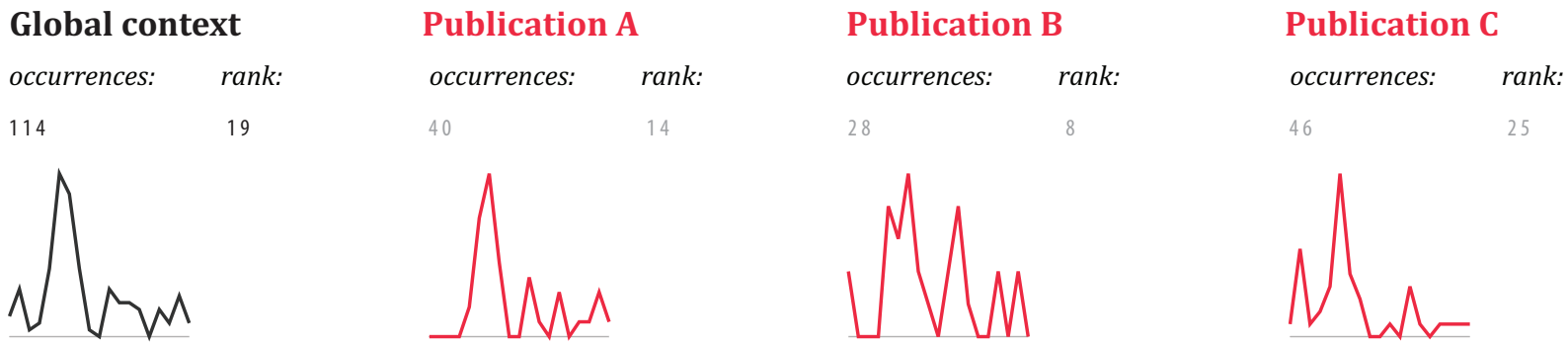


FIGURE 8 – Parallel line graphs used to compare word use across publications

Quantifying perceptions of Islam in the media

A more suitable example of separate line graphs is found in a 2016 work examining how Islam and Muslims were portrayed in the French news 1997–2015.[53] The study employed many methods similar to the ones used in this study, and documented their data scope and methods admirably for fellow practitioners. The interactive "scrollstory" (scrollable story) user interface employs many different visualization types throughout, but the most relevant again is the comparison of word use over time. Unlike the Google Ngram Viewer, showing the line graphs separately avoids the problem of implying correlation between the different line trajectories. As an added bonus, the visualization not only compares the word use over time and across publications, but also provides a common baseline for each by charting the average word use in global news as well. Some significant drawbacks, however, include a lack of axis legend, which in effect leaves upticks in the timeline with no year label for context, as well as a lack of common scale for the separate charts: a line graph showing 28 occurrences appears to reach the same heights as one charting 144 occurrences in the news FIGURE 8.

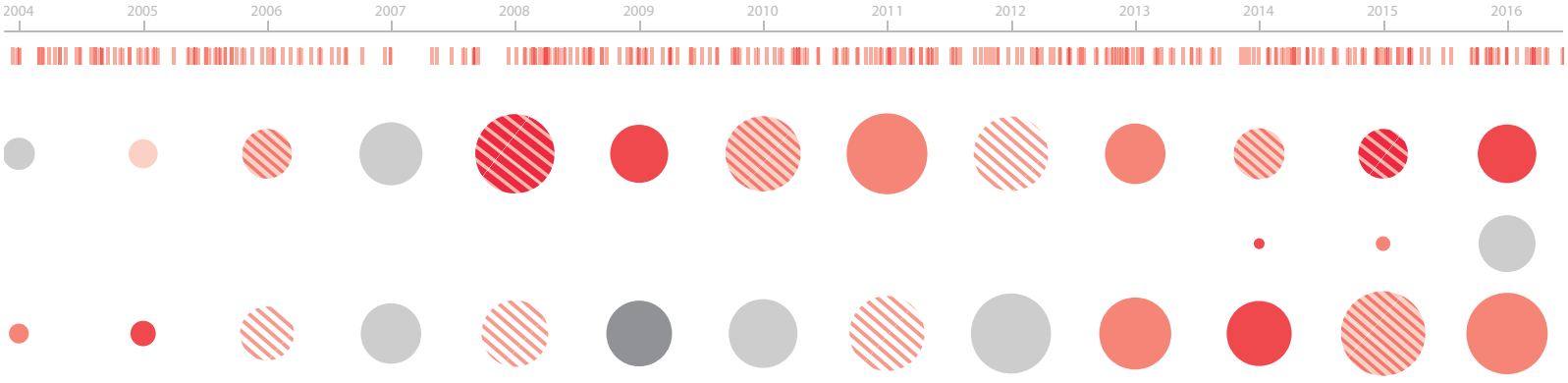
Mo Magazine Dataviz

A particularly versatile news visualization moving away from the use of various line graphs is found in Mo Magazine's online Dataviz page.[54] The deceptively simple user interface combines color-coding for sentiment designation and circle size to indicate the number of articles mentioning a particular topic FIGURE 9. The visualization compares themes across time, side by side, with sentiment data for an added layer of information, and the user has the ability to add a multitude

53 Bellon, "Islam, media subject…"

54 Mondiaal Nieuws. "Mo Dataviz."

FIGURE 9 – Topic coverage across time, as visualized by Mo Magazine



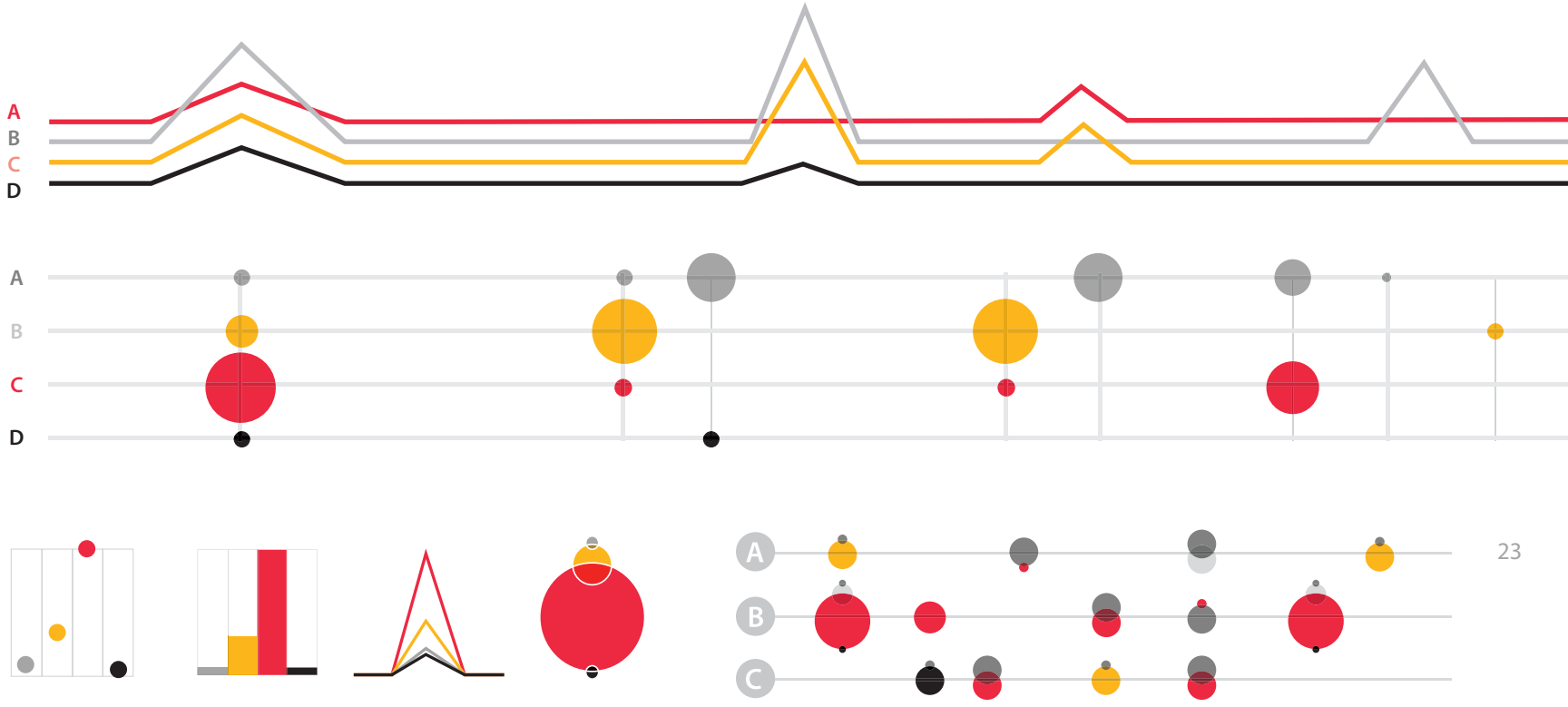
of themes to compare. The themes are pre-defined so the selection is in that sense finite, but the list of themes to choose from is in itself quite extensive, and affords the user a very powerful means of exploring the data.

Unfortunately the dataset used to produce the visualization is not known, as documentation for the visualization seems to be missing at this time or is still in the process of being written. Consequently it is difficult to comment on the nature of the data being visualized, and whether it is done in the most effective manner possible.

5.2 Initial prototype drafts

Initial sketching for the prototype FIGURE 10 touched on many of the same ideas as the example media visualizations outlined previously, using circles of various sizes to denote the number of times a word occurred in a time period, or parallel timelines of line graphs showing the use of words over time, and so on. These provided a general direction for the visualization, primarily the framework of parallel timeline axes for each publication. There were several competing candidates for the basic visual "unit" that would populate the timelines, among them sequential circles, clustered bar charts, peaked line graphs, Venn or Euler diagrams, and various permutations and combinations of these units. The original units measured and averaged the data one day at a time, on a very granular level.

Some of the drawbacks of these drafts were the same that plagued the extant media visualizations described in the previous section. For example, the line graphs lacked a common baseline for accurate comparisons, and seemed to imply some correlation in the data peaks, when in fact any real



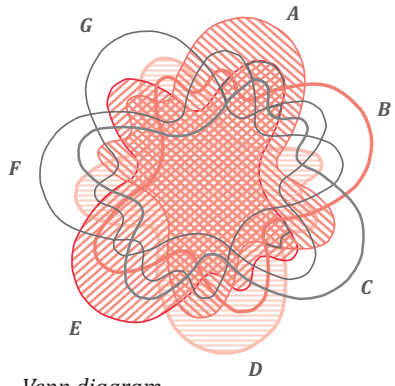
correlation could not realistically be seen from this basic treatment of the dataset. It is quite difficult as well to quantify and compare the area of circles with any certainty,[55] or to distinguish any pattern from groupings of different sized circles. This is true also for Venn and Euler diagrams,[56] though they potentially bring an added layer of information to the visual. In the end, while none of the initial drafts proved successful, it was the introduction of these overlapping diagrams which opened the door for alternate visualization ideas.

FIGURE 10 – Some initial sketches for visualizing words or topics over time and across publications.

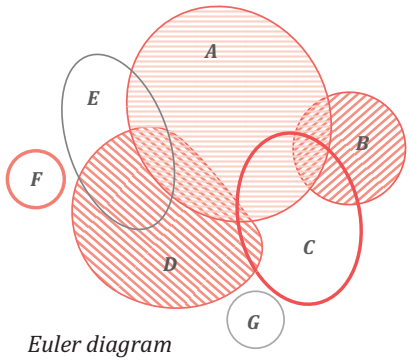
Each color represents the occurrence of a different word or topic, and the individual unit's (circles, Euler diagrams, bar charts) location on the X axis signifies a different chronological point in time.

55 Koponen, Hildén, Vapaasalo, *Tieto Näkyväksi*, 188.

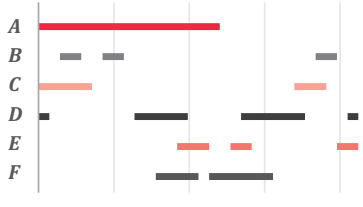
56 Ibid, 227.



Venn diagram



Euler diagram



Linear diagram

5.3 Alternate visualizations

The aforementioned visualizations and drafts touched on the same subject of media and word/topic occurrences in many ways, and many of them employed some of the same basic elements: horizontal timelines, parallel axes to denote separate sets, color coding for topic or sentiment designation, and quantifying words or topics for comparison. However, all failed to address an important aspect of the data: the proportion of articles that included words A *and* B (denoted as A∩B), or B *and* C (B∩C); that is to say, the co-occurrences of words or topics, or different *sets* as it is commonly referred to in the set theory branch of mathematics. Envisaging and communicating co-occurrence of variables does not seem to be common outside mathematics or computer science publications about set notation, and even therein the element of time remains unaddressed.

Traditionally sets, with their various intersections and unions, have been visualized using *Venn* or *Euler diagrams*. The main difference between these two diagrams is that Venn diagrams show all possible outcomes of set unions and intersections, whereas Euler diagrams visualize only the non-empty sets. Neither however, is particularly suited for visually quantifying data, and become very difficult to comprehend when more than three sets are visualized at once.[57] Various versions and improvements have been proposed to these traditional diagrams, but *linear diagrams*, a collection of parallel horizontal lines denoting different set relationships by their vertically overlapping segments have been shown as a more viable option for coherent set notation.[58] Not only are linear diagrams visually less cluttered and thereby easier to produce in an automated fashion, it has been shown in recent empirical studies that they outperform more traditional forms of set notation in both speed of comprehension as well

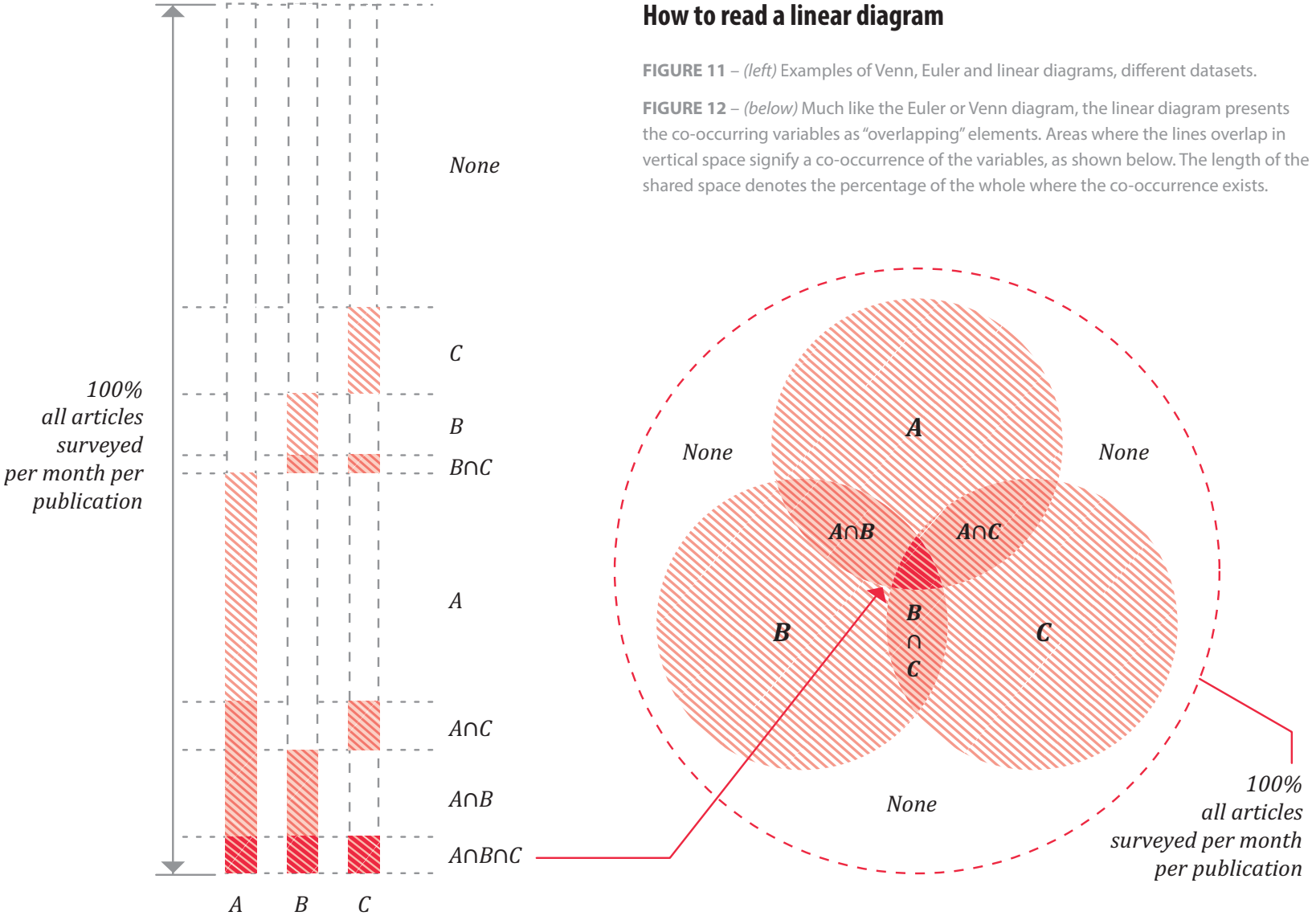
57 Gottfried, Björn. "A Comparative Study of Linear and Region Based Diagrams." Journal of Spatial Information Science, 3-20.

58 Rodgers, Peter, Gem Stapleton, and Peter Chapman. "Visualizing Sets with Linear Diagrams." ACM Transactions on Computer-Human Interaction, 1-39.

How to read a linear diagram

FIGURE 11 – *(left)* Examples of Venn, Euler and linear diagrams, different datasets.

FIGURE 12 – *(below)* Much like the Euler or Venn diagram, the linear diagram presents the co-occurring variables as "overlapping" elements. Areas where the lines overlap in vertical space signify a co-occurrence of the variables, as shown below. The length of the shared space denotes the percentage of the whole where the co-occurrence exists.



as accuracy of viewer assessment.[59] It can furthermore be extrapolated that unlike more traditional diagrams, linear diagrams are much better suited for set comparisons where multiple diagrams are displayed side by side, as line length on a common baseline is much easier to visually quantify than volumes of space.[60]

5.4 Prototype

In this particular application of linear diagrams, some minor changes were undertaken for better integration into a timeline setting. Most notably, it was determined that flipping the conventional horizontal orientation into a vertical configuration better supported comparisons of one set visualization to the next on a horizontal timeline. Secondly, though color-coding has been shown to be generally irrelevant in individual linear diagram applications,[61] color proved to be a necessary labeling tool in this setting, as the legend could not be reasonably repeated *ad infinitum* across multiple axes. It was furthermore decided that the number of sets (i.e. columns of lines, keywords) should be limited to three at a time, since even linear diagrams lose some of their comprehensibility when too many sets are presented side by side at one time.[62-63] Furthermore, for the purposes of simple media variable comparison more sets would not necessarily have been appropriate.

The procedure by which the visualization creates the linear diagrams is a repetitive checking process that is iterated for each month of the year, and individually for each publication to produce unique linear diagrams for each month of each publication. A single iteration of the process involves checking each article in the chosen time period for the occurrence of the keywords — preferably a word-stem such as *migr,* which would encompass in its results all derivative words such as *migrant(s)*, *migration, immigrant(s), immigration,* etc. (more on this in the discussion of possible *Extensions and Further Development*). The query returns a true/false response for each keyword, which then is translated into a corresponding length of line as a percentage of all articles surveyed. Since the upper sets (not A, not A or B... etc.) produce no visible result, the linear diagrams communicate more the relative nature of the results rather than concrete percentages. The sample size for each publication and each unit of time nevertheless remains the same (250, 500, 750... see restrictions of the API in the *Materials and Methods* chapter). Since the results communicated by the linear diagram are relative rather than concrete, a vertical axis legend was also deemed superfluous and was eliminated to minimize clutter.

The end result is a prototype where each individual bar denotes a different keyword/topic and is accordingly color-coded to separate it from its neighbors. The linear diagrams

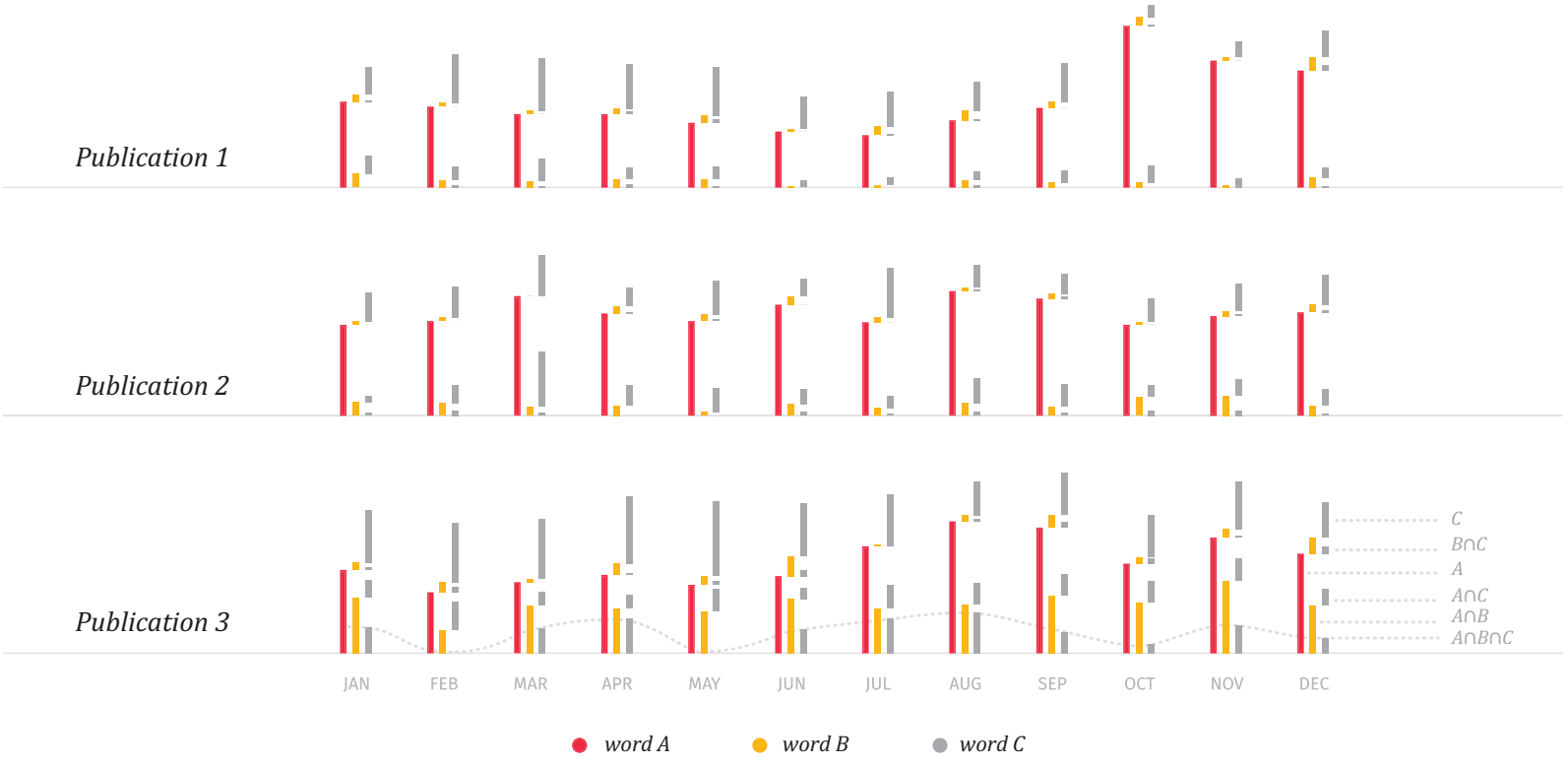
59 Chapman P., Stapleton G., Rodgers P., Micallef L., Blake A., "Visualizing Sets: An Empirical Comparison of Diagram Types" in *Lecture Notes in Computer Science*.

60 Koponen, Hildén, Vapaasalo, *Tieto Näkyväksi*, 187.

61 Luz, Saturnino and Masood Masoodian. "A Comparison of Linear and Mosaic Diagrams for Set Visualization." Information Visualization.

62 Ibid.

63 Caleydo UpSet tool (http://caleydo.org/tools/upset/) uses linear diagrams on greater sets, but the timeline axis is missing and the number of sets displayed is so large that a comparison over time would be impossible.



are grouped in sets of three keywords per month, sequentially spaced left to right along the timeline, with each individual set representing a single month of data. The individual linear diagrams, when placed side by side on this horizontal time axis, can be visually compared to chart changes in co-occurrence patterns. A simple example of the prototype is presented with labels in FIGURE 13. The following pages include further examples of the linear diagrams in use on a temporal scale, illustrating varying topics of interest, with discussion on patterns and possible insights.

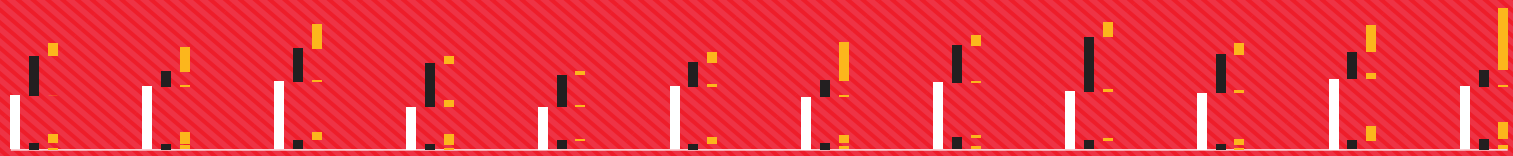
FIGURE 13 – A simple example of the prototype. The third publication has been labeled to highlight the various presented sets, as well as pattern changes in the co-occurrence of words A, B, and C.

2015

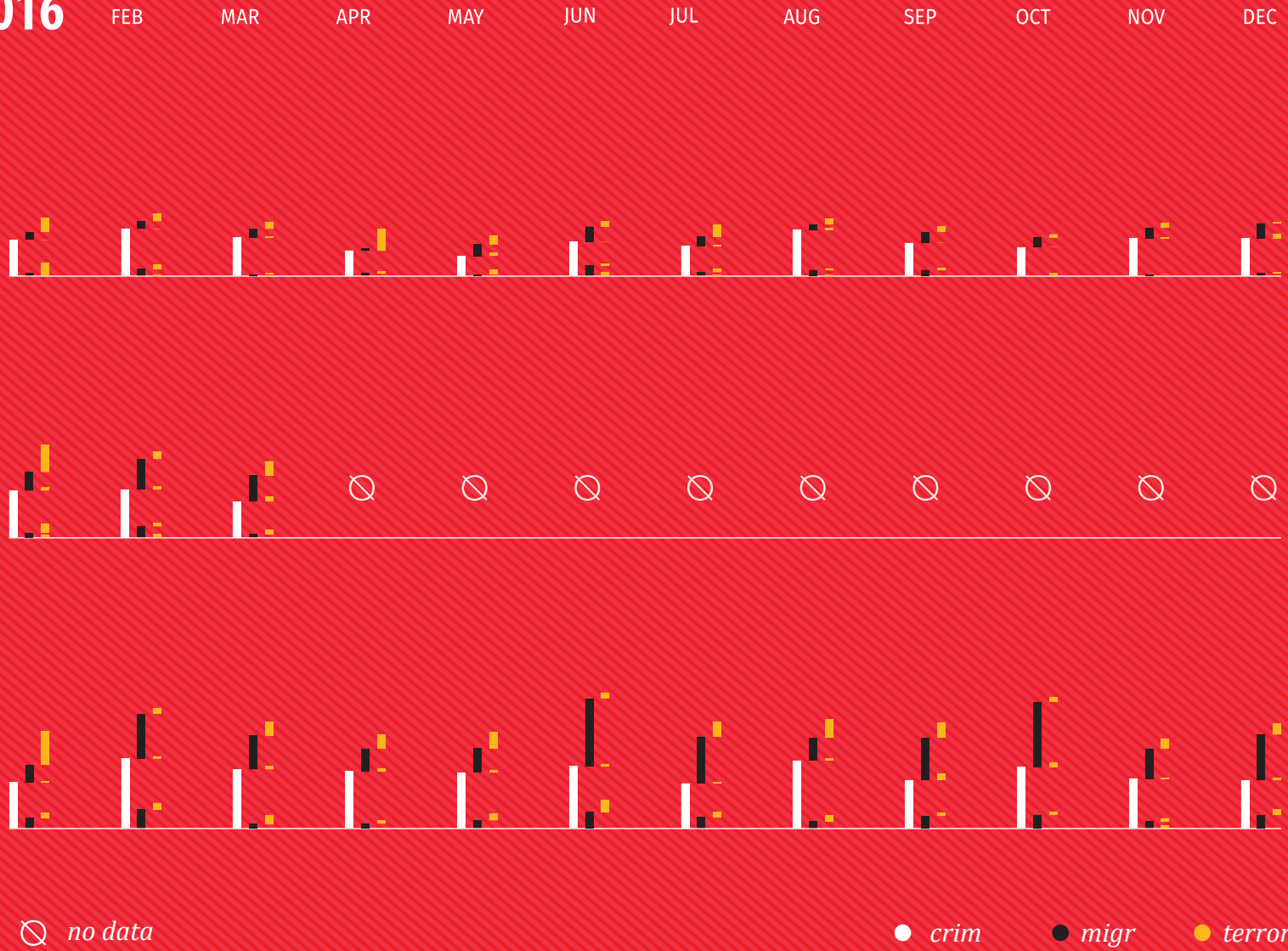
FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

The Daily Mail

The Telegraph



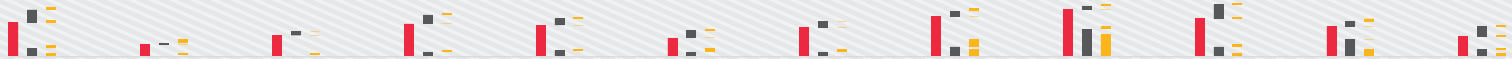
The Guardian


2016
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
no data
crim
migr
terror

2015

FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

The Daily Mail



The Telegraph

The Guardian


2016
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
no data
migr
refug
asylum

2013

FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

The Daily Mail

The Telegraph

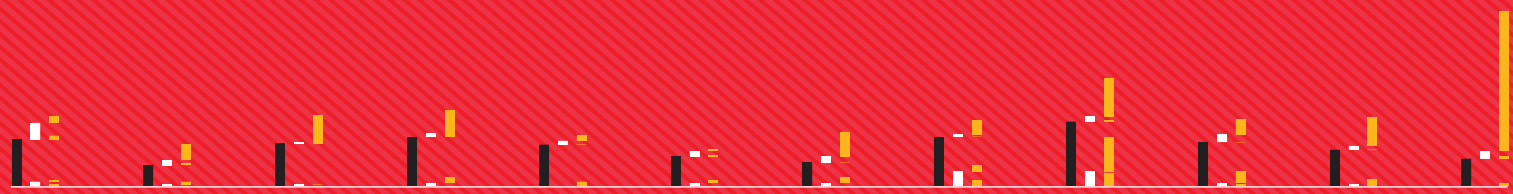
The Guardian


2014
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
migr
africa
syria

2015 FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

The Daily Mail

The Telegraph

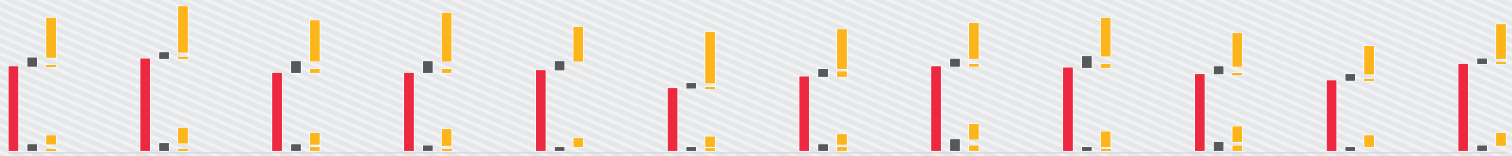


The Guardian

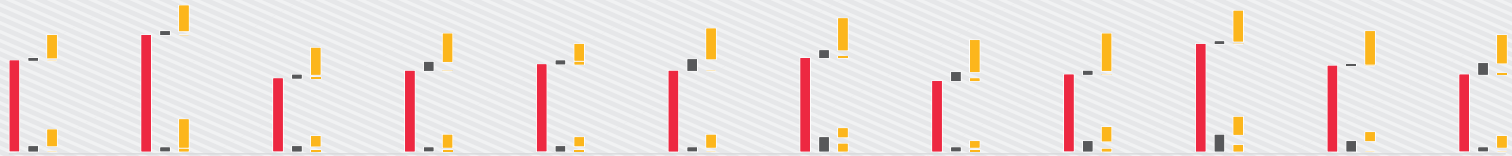

2016
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
no data
migr
africa
syria

2011

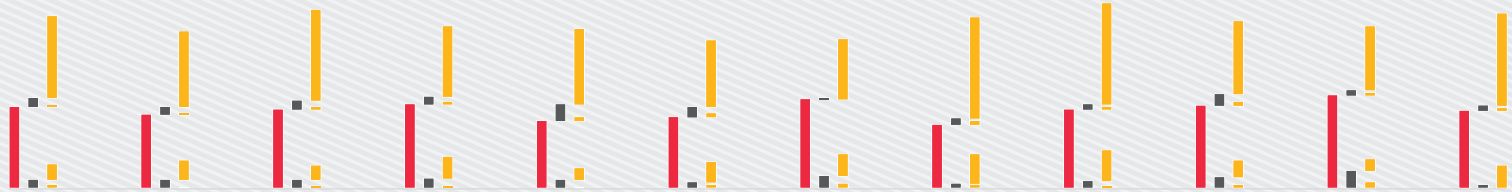
FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC



The Daily Mail



The Telegraph

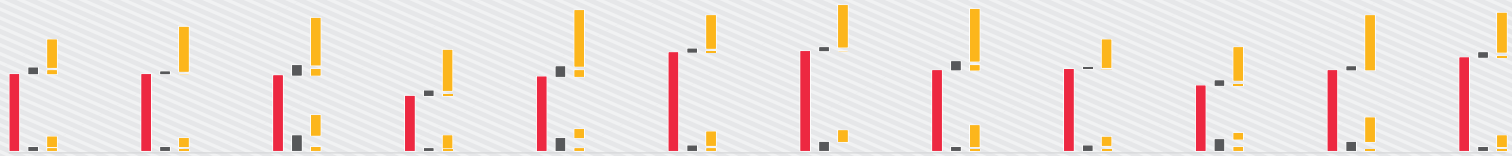


The Guardian

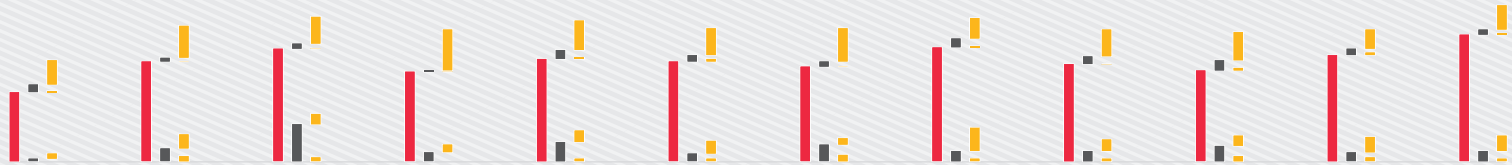

2012
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
eu
migr
crim

2013

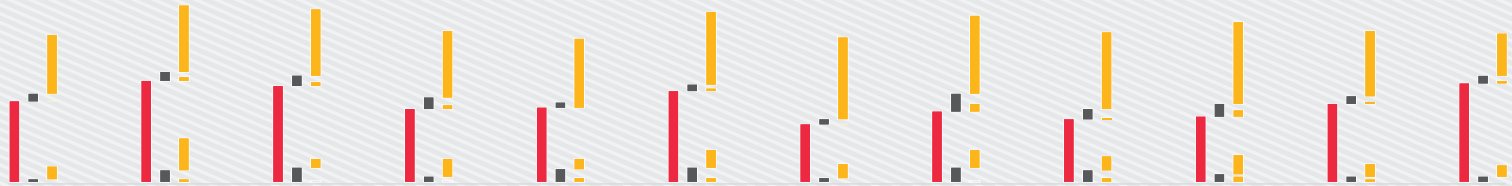
FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC



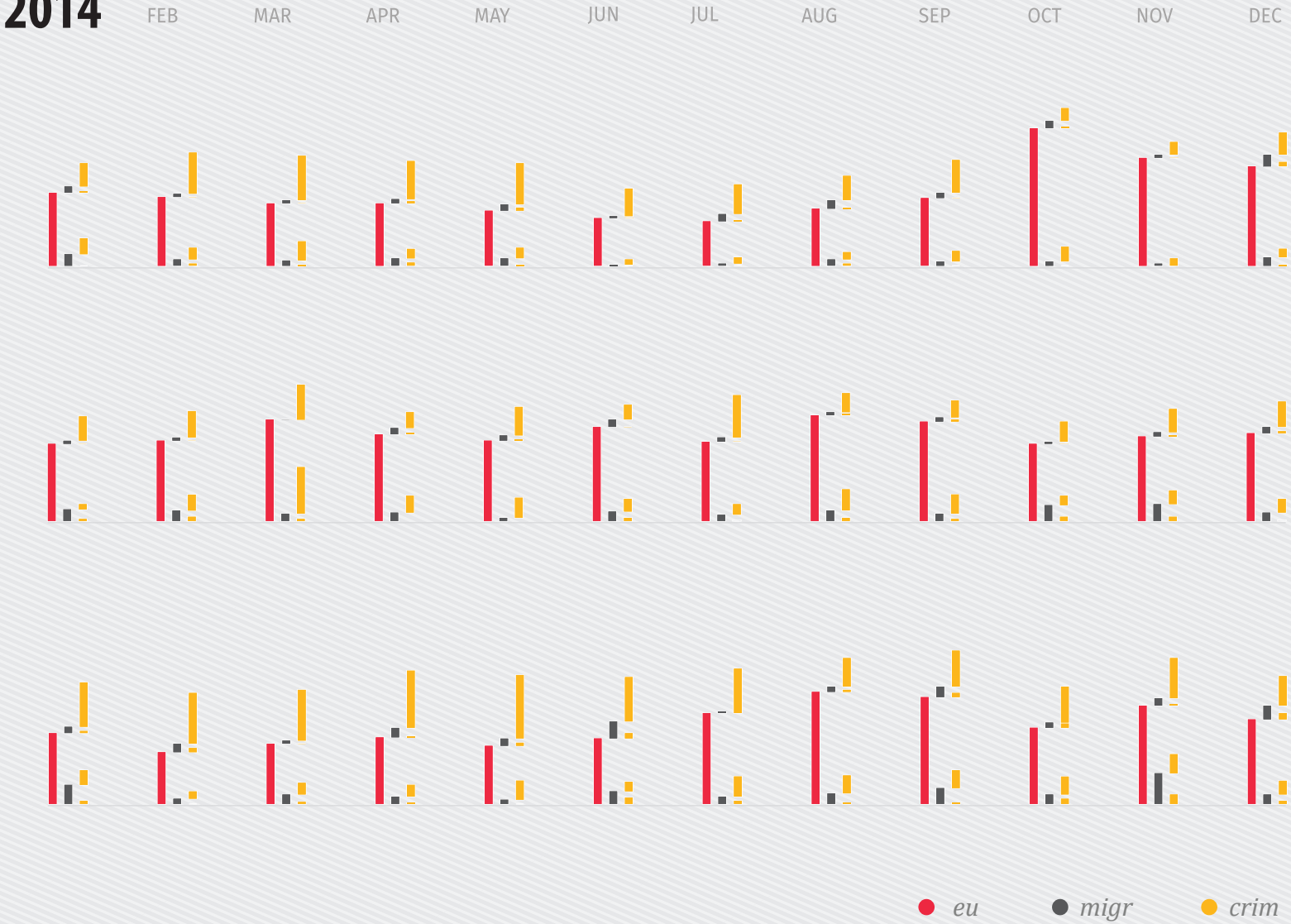
The Daily Mail



The Telegraph

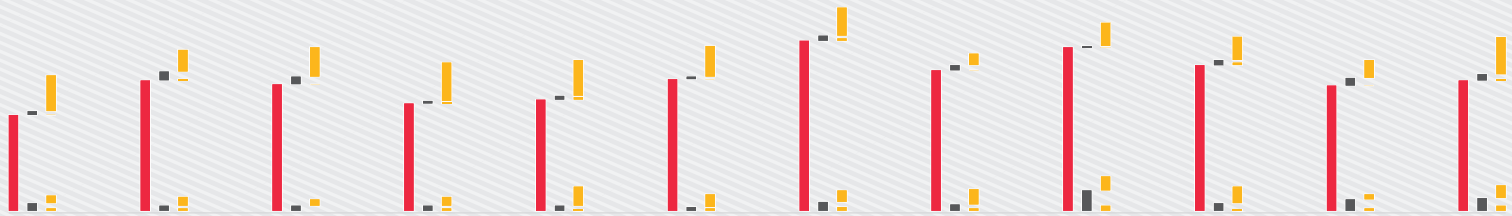


The Guardian

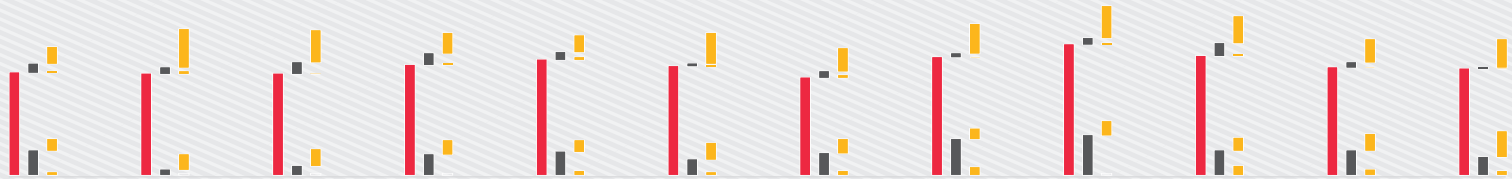

2014
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
eu
migr
crim

2015

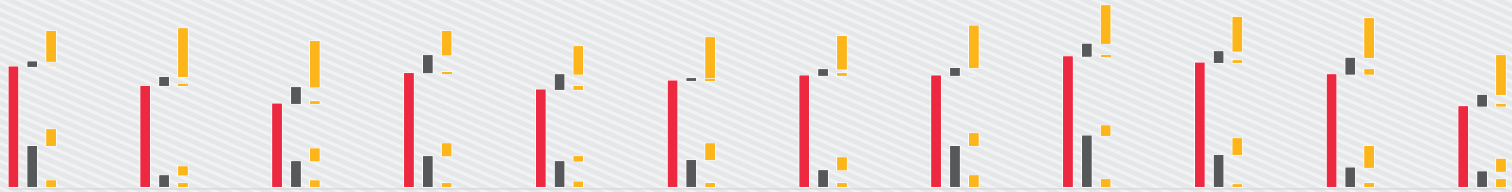
FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC



The Daily Mail



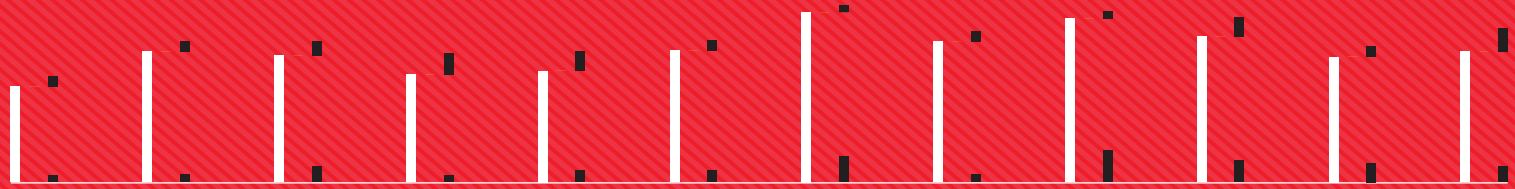
The Telegraph



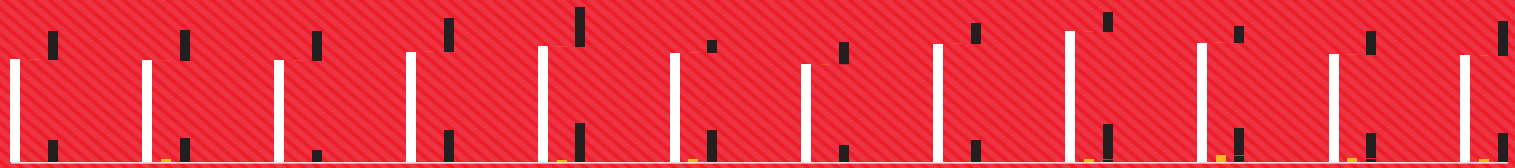
The Guardian


2016
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
no data
eu
migr
crim

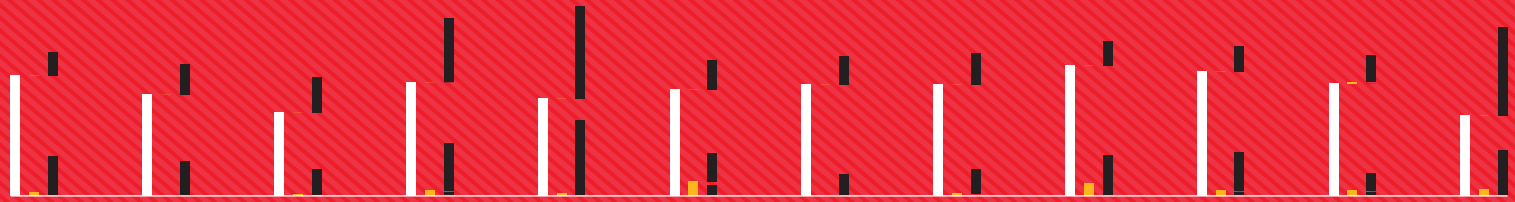
2015 FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC



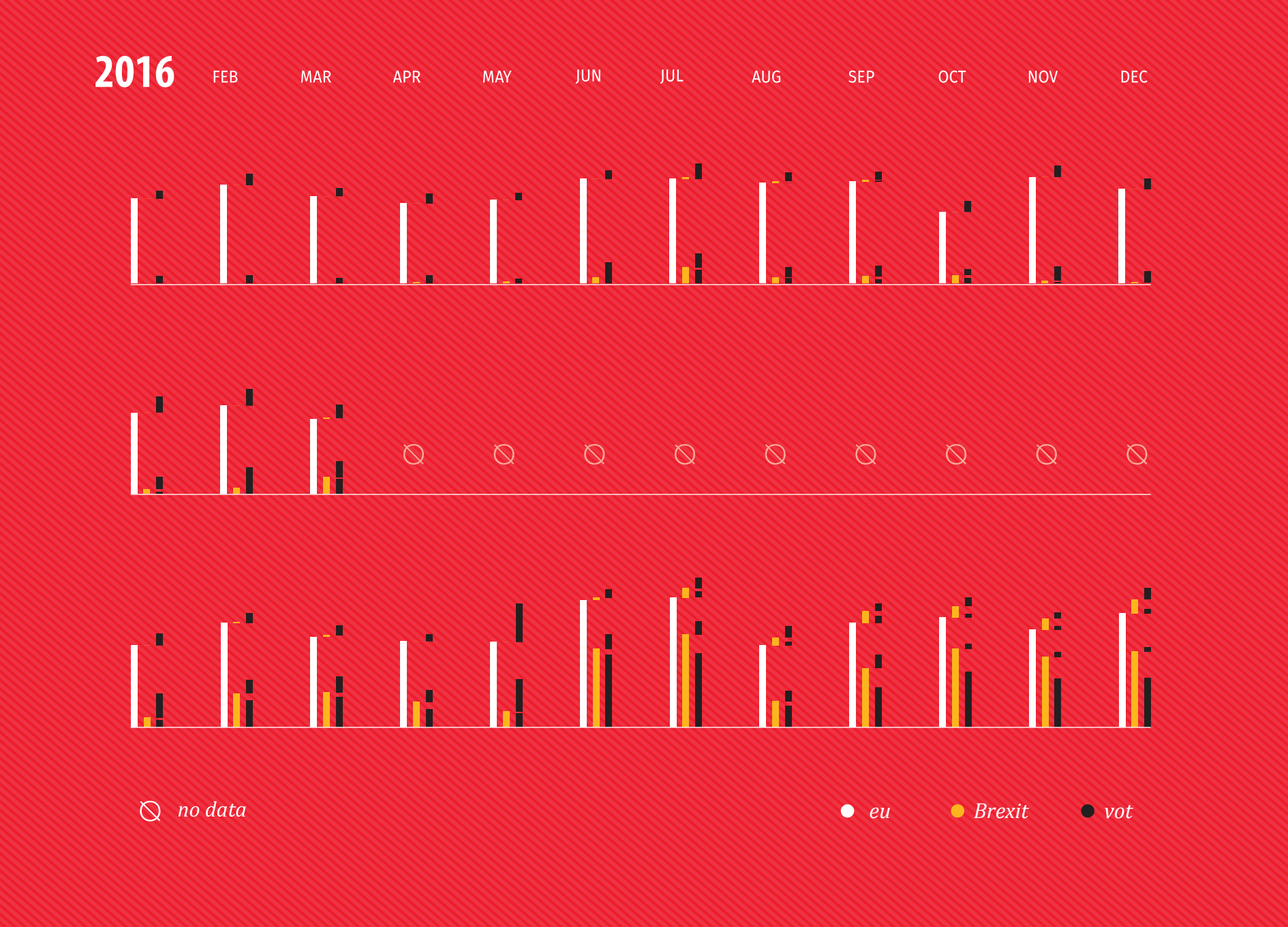
The Daily Mail



The Telegraph



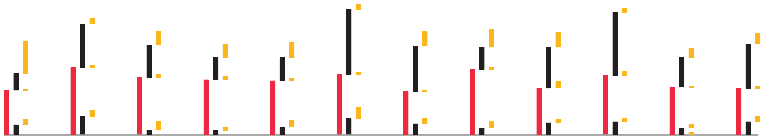
The Guardian


2016
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
DEC
no data
eu
Brexit
vot

6. Case study

It is evident from the sample visualizations that certain patterns are visible within the data, and one might be tempted to make conclusions based on the patterns therein. It should again be emphasized, however, that the co-occurrence of words or topics does not equal article sentiment; nor can it be taken as a sign of media bias or even evidence of prevailing public opinion. All that co-occurrence of keywords signifies is the rate of co-occurrence, which by itself without temporal comparisons and further in-depth analysis of the underlying data is relatively meaningless. It could however highlight areas on a timeline which would be suitable targets of further investigations.

The sorts of observations that can be made from these visualizations are primarily relative in nature, though they could be used to bring forth questions and working hypotheses for further analysis of the underlying data. What follows is a brief discussion of some of the timeline samples laid out in the previous pages, with relative observations and purely speculative questions that might arise from each particular visualization set.



crim*/*migr*/terror*

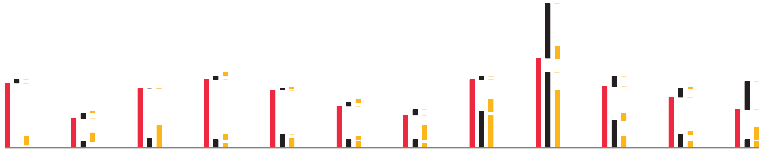
Possible derivative words/topics: *crime, criminal(s), criminology, immigrant(s), migrant(s), immigration, migration, terror, terrorism, terrorist(s)*

This initial set was chosen to examine whether news media were reporting on immigration often in conjunction with topics of crime or terrorism, as some of the conservative rhetoric exhibited throughout Europe at this time seemed to suggest that increased immigration into the country brought with it higher crime rates and higher likelihood of terrorist attacks. The question in this, of course, is not whether we can determine from this dataset that immigration and crime coincide in the real world (one might conclude this if the news reported on the topics together), nor can we conclude that the news media was trying to coerce public opinion in that direction, contrary to the truth… We can simply glean whether or not the news was reporting on these topics together.

Observations: While there are definite trends in time of immigration being mentioned in conjunction with crime, thankfully it is not a large trend, nor does immigration get mentioned in conjunction with terror in any large number of articles, it would seem. On the other hand, in 2015 the report-

ing on crime-related topics does not in itself vary much over time, but there is an uptick in unrelated articles discussing migration and terror. It is also interesting that *The Guardian* reported on migration- and terror-related subjects much more than the other two sources, and this is found throughout the dataset on several topics of note such as crime, EU, employment etc.

Questions: Why does *The Guardian* exhibit such high percentages of hot topic issues compared to the other publications? Is it simply a tendency toward controversial topics? Or could it be taken as a sign of sensationalism? Maybe it is conversely a sign of a more news-oriented publication?



migr/refug*/asylum*

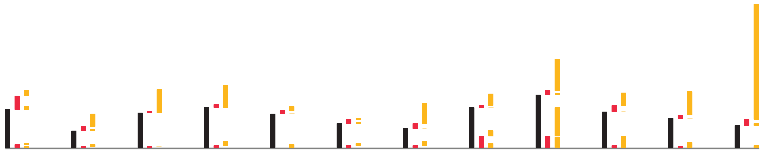
Possible derivative words/topics: *immigrant(s), migrant(s), immigration, migration, refugee(s), refuge, asylum, asylum seeker*

Comparing these three word-stems is particularly interesting, as it can shed light on what terms were used and in what context when referring to a certain subset of people in news coverage. The words used to talk about people in an immigration context are very telling in themselves; these three were some of the most neutral ones identified in the European Migration Network's *Asylum and Migration Glossary 3.0.*[64] Some more conspicuously biased terms included "illegal" or "undocumented alien," among others. Fortunately it seems the news media at least within this dataset chose to avoid such contentious terminology.

Observations: It is apparent that something newsworthy occurred in September of 2015 that caused increased discourse on all three topics of immigrants, refugees and asylum seekers, with a fairly high rate of co-occurrence between the terms. Whatever the news event causing the increased coverage, the co-occurrence might suggest that immigrants were being referred to primarily as refugees and asylum seekers at this time. At the next uptick in conversation in 2016 however, the immigration topic had barely any overlap with discussion of refugees and asylum seekers, in fact discussion on the latter two topics had all but ceased.

Questions: Does this change in topic co-occurrence indicate that the sort of immigration being discussed around the Brexit vote in June of 2016 just happened to only address immigrants who were *not* refugees or asylum seekers, or was there a subliminal shift in narrative which unintentionally or otherwise downplayed the humanitarian needs of immigrants in favor of xenophobic rhetoric? Or perhaps it might indicate a shift from discussion of refugees to one of immigrant Europeans living in Britain at the time?

64 European Commission. "Asylum and Migration Glossary 3.0." European Website on Integration: Migrant Integration Information and Good Practices.



migr/africa*/syria*

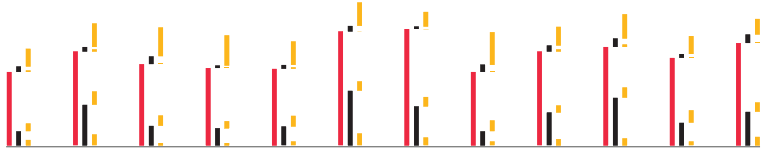
Possible derivative words/topics: *immigrant(s), migrant(s), immigration, migration, African(s), Africa, Syrian(s), Syria*

The two keywords referring to Africa and Syria were chosen because purportedly much of the migration during this time period originated from those two geographical regions, albeit one is very broad and immigration could of course have been discussed in the context of specific African countries as well.

The co-occurrence of these three words might well be used to track patterns of migration itself, or at least when the news media felt it pertinent to cover the topic. It would be particularly interesting to compare the peaks in topics of discussions with the actual migration statistics, to see what kind of correlation there actually is between news coverage of migration and migration itself.

Observations: News coverage of conflicts in Syria, and news about immigration from Syria, can debatably be distinguished by the occurrence of the Syrian variable on its own with no mention of migration/immigration. Several different peaks in the narrative can be distinguished; some that do largely mention immigration and some that do not.

Questions: It is interesting to note that in 2016 immigration was rarely discussed concurrently with either of the two other topics. Does this mean immigration from those sources diminished greatly? Or that immigration from other countries became more prevalent? Or perhaps immigration was discussed in the media in a more general sense, not referring to any particular nationality or geographic origin?



eu*/*migr*/crim*

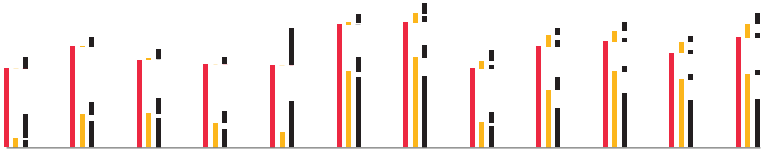
Possible derivative words/topics: *EU, European Union, Europe, European(s), euro(€), immigrant(s), migrant(s), immigration, migration, crime, criminal(s), criminology, (possible unintentionals) euthanasia, eulogy, euphoria…*

These keywords were chosen, much like the *crim/migr/terror* set, because it seemed some prevalent rhetoric leading up to Brexit votes was suggesting that membership in the EU was resulting in increased migration, and thereby increased crime.

Observations: In 2010 reporting on crime far outpaced both Europe-related topics as well as migration, but by 2016 crime was overtaken by the other two. While it can't necessarily be observed in the dataset that all three coincided in news leading up to the Brexit vote, there is a marked increase

in mentions of European topics with migration, and European topics with crime, separately. It is further interesting that while reporting on EU and crime *together* increased from 2010 to 2016, the percentage of articles about crime altogether decreased.

Questions: Did crime rates decrease during this time, leading to an overall decrease in crime reporting? Or was crime being intentionally linked to the EU? Or is the decrease in crime reporting an example of an arbitrary course alteration in public narrative, a shifting away from news coverage that is deemed too violent or depressing?



eu*/brexit/vot*

Possible derivative words/topics: *EU, European Union, Europe, European(s), euro(€), Brexit, vote, voting, voter, (possible unintentionals) euthanasia, eulogy, euphoria…*

Debatably the most interesting narrative patterns in these examples are found in the eu/brexit/vote sample, perhaps because of the very time-bound nature of the term "Brexit "in social discourse. It is also one of the few search terms used herein which can be definitively tied to only one topic variation.

Observations: The pattern quite clearly displays the first emergence of the term *Brexit*, surfacing surprisingly (or perhaps not) in reporting by *The Guardian* in 2015, and quickly picking up steam to become a major topic of discussion in 2016. By the time of the Brexit vote in June of 2016, relatively few articles mention Europe, EU, European(s) without mentioning Brexit or the vote. It would seem that Brexit was also always covered *with* the term vote, so it could be stipulated that from the outset it was reported on as an active voting option, rather than a passive suggestion.

Questions: Do the peaks in the *vote* keyword in 2015 coincide with polling days in Britain? Were there no other votes of note in 2016, or was it just not covered in the news without mentioning the upcoming, or recently past, Brexit vote?

7. Discussion

7.1 Extensions and further development

As an experimental portion of this case study, NodeBox was also used to create an interactive version of this static prototype. Added functionality namely included the ability to change the date ranges and input custom keywords to further examine narrative trends of interest. This ability to interact with the dataset greatly increases the visualization's potential as a viable media analysis tool. Alas, the dynamic version leaves much to be desired at the moment, and further extensions are necessary to produce a truly worthwhile result.

One particular improvement that could be implemented is the optimization of the database-API-visualization pipeline, as the current configuration cannot make use of the full extent of the data, nor is the user interface nearly as nimble and responsive as could be desired. The dataset could be further pre-processed and the API optimized so that the full text articles are not being directly input into NodeBox. Further, some natural language processing (NLTK) tools could be implemented to allow for word stemming of search terms —this could maybe also be part of the pre-processing— so that the user could freely search for any keyword without stemming the word themselves (e.g., "criminal" vs. "crim").

An additional feature that could ideally be integrated in the visualization is bifocal zoom,[65] in other words, showing more detail in one area of the visualization while keeping others in focus. "Zooming in" by changing the averaging unit from months to weeks, or even days, would enable the viewer to explore the data on a more granular level, possibly even leaving some month sets in focus on the same axis to achieve a consistent scale. An example of this bifocal zoom is shown in Figure 14. From there, being able to examine individual articles 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 narrative. The Mo Dataviz visualization[66] achieves this in a particularly novel way, with mouseovers of the secondary timeline revealing previews of the article titles. With the image URL already integrated into each article entry, a visual preview might also add to the user experience. The visualization could also draw on live data instead of a static dataset, for example by connecting it to one of the pre-existing open data news APIs. The result would be a live feed of media narratives that could track changes as they occur.

The visualization could also be deployed in a separate animated form to further underscore differences in patterns over time. For

65 Spence, Robert, and Mark Apperley. "The Encyclopedia of Human-Computer Interaction: Bifocal Display." Interaction Design Foundation.

66 Mondiaal Nieuws, "Mo Dataviz."

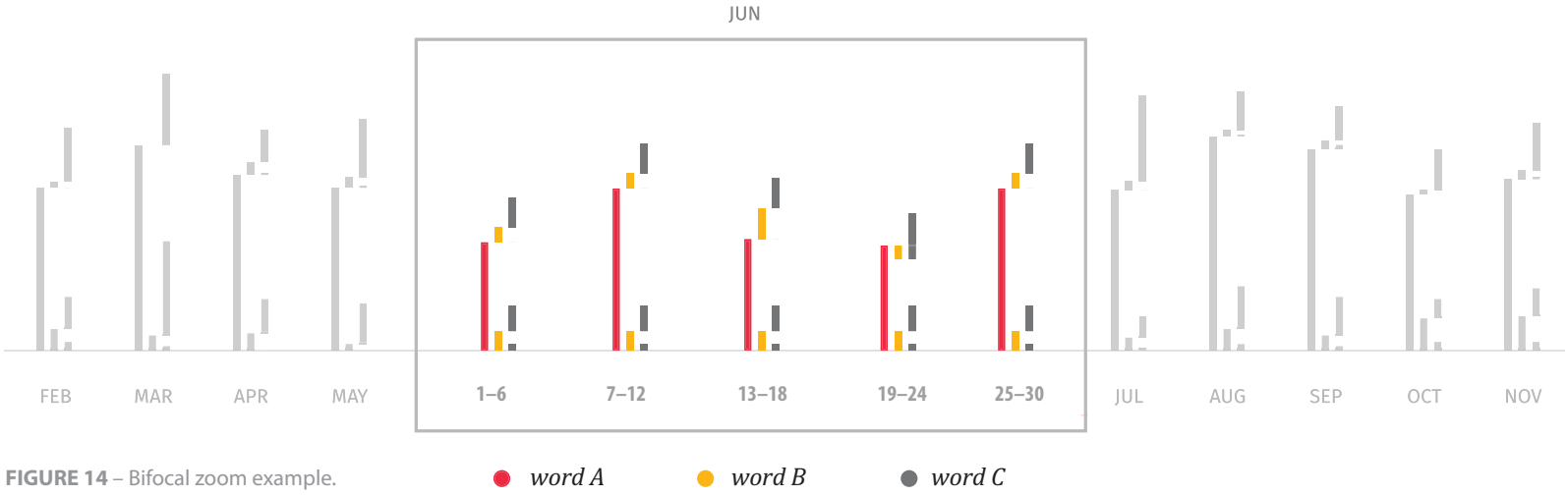


FIGURE 14 – Bifocal zoom example.

example, if the timelines were to scroll horizontally across a display the differences in line heights could become particularly apparent to the viewer. Another layer of information (e.g., notable events covered in the media) could moreover be superimposed on the animation to explain certain marked peaks or trends in the visualization. Adding more levels of detail and the ability to examine the media events behind the trends would be ideal as an extension of the visualization, regardless of its form.

The extension opportunities are not limitless, but needless to say there are many changes that could be undertaken to further enhance this visualization in its capacity as a media analysis tool. Nevertheless, even in its static form it performs its basic necessary functions of setting out the data and providing room for comparisons.

7.2 Reflections on the design process

Strictly speaking, the product of the design process (i.e. the prototype, the visualization) came together very organically from the various pieces of data and gleaned information, and needed very little coaxing in the end to assume its final form. Creating the visual in essence, proved to be the least challenging part of the entire design process.

Though discussion of this process in itself could be limited to those tasks dealing with the prototype's visual aspects, the entire project story arch seems crucial in attaining the end result with any degree of scientific accuracy and data integrity. One particular insight from the

process has been how important this data integrity really is to a data-driven project. The data gathering process, and subsequent visualizations of the dataset, involved an endless series of decisions that could jeopardize the reliability and consistency of the result. Only by undergoing this gauntlet of data-compromising decisions can one arguably come to understand and treat the dataset with the respect it deserves, a crucial characteristic in an accountable data designer —or for that matter, any journalist or researcher.

Consequently, a desire to be truly accountable for each decision in the process also made conducting a form of "literature review" of existing media visualizations somewhat difficult. Finding current visualizations of media through online searches seemed unnecessarily haphazard, relying heavily on what the search engine of choice deemed worthy. In this sense the results may have been influenced by general popularity of the sites, or even ranking based on paid advertising. Finding visualizations in printed books was also rather arbitrary unless the books specifically dealt with media studies, in those cases the forms of visualization were also more traditional and ubiquitous.

The most innovative and interesting examples of media visualizations were found through happenstance of browsing search results and online data visualization blogs or periodicals, but as discussed in the *Review of media visualizations* section, each had their own drawbacks as a truly representative media narrative visualization. Furthermore unlike academic papers, the sorts of visualizations that abound online today rarely cite sources or enumerate design principles used therein, with some few commendable exceptions.[67]

The academic side of the literature review (namely outlined in the *Alternate visualizations* section) of course proved much simpler thanks to copious citations, references and concrete result metrics, and provided the necessary backbone for the working prototype that could not be found from popular forms of media visualization. The process, among many other things, instilled a deep appreciation for the academic literature surrounding visualization types, and good documentation in general. Especially the exhaustive documentation of various emerging technologies online proved to be invaluable in this process, and those lacking in documentation proved to be veritable stumbling blocks along the way.

Despite some documentation setbacks and a rather high learning curve for several of these emerging technologies (especially from a traditional graphic designer's perspective), the results of this process are extremely gratifying. Keeping course in pursuing the original goal, not to mention achieving said goal, and even helping to bring something novel into the field have made for an extremely rewarding experience.

67 Bellon, "Islam, media subject…"

8. Conclusion

While this study was in many ways an exploration of a particular experimental process, and the technical means by which to produce a desired visual prototype, the end product in itself yields unexpectedly meaningful results. Arguably a visualization comparing sets of concurring variables over time has not previously been presented in the field of data visualization, with or without several parallel data sources or the intent of identifying patterns in discourse. Though at best these novel visuals can lead to a series of relative observations and hypotheses about the underlying causes of patterns, it can most definitely be concluded that the prototype is a successful means of utilizing data visualization to examine differences in media narrative patterns over time and across publications.

The prototype itself of course presents the co-occurrence of these variables in vertical linear diagrams sequentially on parallel timelines. Using such frugal visual elements, combined with only a minimal amount of textual analysis, is perhaps as close to a simple reporting of the raw data as possible. It also seems to be a reasonably effective tool for identifying wrinkles in the narrative social fabric (i.e. patterns in discourse). Of course further studies are necessary to determine the specific extent of effectiveness, and whether other new forms of visualization might not perform better in speed and accuracy of viewer assessment.

The applications nevertheless could be manifold, and could especially be instrumental in functions such as media trend studies and in pursuits of journalistic accountability and oversight. Pursuing accountability is particularly relevant now more than ever, as real and fake news alike exercise power in the public sphere, deciding what society experiences and how it is experienced. The media we as designers and journalists generate undoubtedly both reflects and perpetuates public opinion as a primary source of evidence and a forum for ideas, opinions, and influence,[68] and therefore we have all the more responsibility to be mindful of the narratives we disseminate.

68 McQuail, Denis. M*cquail's Mass Communication Theory,* 4.

9. References

- Bellon, Pierre. "Islam, media subject: How to quantify the perception of Islam in the media." Data Driven Journalism. Last modified 5.2. 2017, accessed 7.2.2018. *http://datadrivenjournalism.net/featured_projects/islam_media_subject_how_to_quantify_the_perception_of_islam_in_the_media*.

- Chapman P., Stapleton G., Rodgers P., Micallef L., Blake A. "Visualizing Sets: An Empirical Comparison of Diagram Types". In: Dwyer T., Purchase H., Delaney A. (eds) Diagrammatic Representation and Inference. Diagrams 2014. *Lecture Notes in Computer Science*, vol 8578. Springer, Berlin, Heidelberg, 2014.

- Court of Justice of the European Union (CJEU). *Ryanair Ltd v PR Aviation BV.* Last modified 15.1.2015, accessed 31.1.2018. *http://curia.europa.eu/juris/document/document.jsf?docid=161388&doclang=EN*.

- European Commission. "Asylum and Migration Glossary 3.0." European Website on Integration: Migrant Integration Information and Good Practices. Last modified 31.10.2014, accessed 6.12.2016. *https://ec.europa.eu/migrant-integration/librarydoc/asylum-and-migration-glossary-30*.

- European Commission. "Commission proposes copyright exception for researchers." European Commission: Policies, Information and Services. Last modified 14.9.2016, accessed 7.2.2018. *http://ec.europa.eu/research/index.cfm?&na=na-140916&pg=newsalert&year=2016*.

- Google. "Google Books Ngram Viewer." Accessed 6.3.2018. *https://books.google.com/ngrams*.

- Gottfried, Björn. "A Comparative Study of Linear and Region Based Diagrams." *Journal of Spatial Information Science* 2015 (10): 3-20. doi:10.5311/JOSIS.2015.10.187. *https://doaj.org/article/ 0513c6dd067148b1b525bdfa4cacf3ef*.

- *The Guardian*. "Terms of Service." Accessed 31.1.2018. *https://www.theguardian.com/help/terms-of-service*.

- Hill, Charles A. and Marguerite H. Helmers. 2004. *Defining Visual Rhetorics*. Mahwah, NJ: Lawrence Erlbaum.

- Internet Law Treatise. "Fair use." Accessed 19.2.2018. *https://ilt.eff.org/index.php/Copyright:_Fair_Use*.

- Internet Law Treatise. "Trespass to Chattels." Accessed 19.2.2018. *https://ilt.eff.org/index.php?title=Trespass_to_Chattels&redirect=no*.

- Jensen, Klaus Bruhn. *A Handbook of Media and Communication Research*. 2nd ed. ed. GB: Routledge Ltd - M.U.A. 2012, 109-110, 262-263.

- Julkisen Sanan Neuvosto. "Journalistin Ohjeet Ja Liite." Accessed 12.9.2017. *http://www.jsn.fi/journalistin_ohjeet*.

- Jyrkiäinen, Jyrki. “Media landscapes: Finland.” European Journalism Centre. Accessed 31.1.2018. *https://medialandscapes.org/country/finland*.
- Kerr, Orin S. “Norms of Computer Trespass.” SSRN. Published 3.5.2015. Accessed 19.2.2018. *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2601707*.
- Kopiosto. “Copying in schools and educational institutions: What the license does not permit.” Kopiosto Copyright Society. Accessed 21.3.2018. *http://www.kopiosto.fi/kopiosto/copying/schools/en_GB/licence_does_not_permit/*.
- Kopiosto. “Copying in schools and educational institutions: Scope of permissible copying in teaching.” Kopiosto Copyright Society. Accessed 21.3.2018. *http://www.kopiosto.fi/kopiosto/copying/schools/en_GB/scope_of_permissible_copying*.
- Koponen, Juuso, Jonatan Hildén, Tapio vapaasalo. *Tieto Näkyväksi*. Helsinki: Aalto ARTS Books, 2016, 185-193, 227.
- Krippendorff, Klaus. *Content Analysis : An Introduction to its Methodology*. United States: University of Michigan, 2004.
- Lee, Timothy B. “Court rejects LinkedIn claim that unauthorized scraping is hacking.” Ars Technica. Published 15.8.2017. Accessed 31.1.2018. *https://arstechnica.com/tech-policy/2017/08/court-rejects-linkedin-claim-that-unauthorized-scraping-is-hacking*.
- Luz, Saturnino and Masood Masoodian. “A Comparison of Linear and Mosaic Diagrams for Set Visualization.” *Information Visualization*: 1473871618754343, 2018. doi:10.1177/ 1473871618754343. *https://doi.org/10.1177/1473871618754343*.
- McLean, Susan and Mercedes Samavi. “Data for the taking: using website terms and conditions to combat web scraping.” Lexology. Published 12.3.2015. Accessed 31.1.2018. *https://www.lexology.com/library/detail.aspx?g=5f0b65dd-f699-4c81-a1bc-53e1cebb1d34*.
- McQuail, Denis. *Mcquail’s Mass Communication Theory*. 6th ed. London ; Thousand Oaks, Calif.; Thousand Oaks, CA: Sage Publications; SAGE Publications, 2010, 4-357.
- Mondiaal Nieuws. “Mo Dataviz.” Accessed 6.3.2018. *http://dataviz.mo.be*.
- Muratovski, Gjoko. 2016. *Research for Designers: A Guide to Methods and Practice*. London: Sage Publications.
- National Readership Survey. “Newsbrands: Print/PC.” Accessed 6.3.2018. *http://www.nrs.co.uk/latest-results/nrs-padd-results/newspapers-nrspaddresults*.
- Nielsen, Rasmus Kleis. “Where do people get their news? The British media landscape in 5 charts.” *Medium.* Published 30.5.2017. Accessed 6.3.2018. *https://medium.com/oxford-university/where-do-people-get-their-news-8e850a0dea03*.
- OfCom. “News consumption in the UK.” Published 24.3.2015. Accessed 6.3.2018. *https://www.*

ofcom.org.uk/research-and-data/tv-radio-and-on-demand/news-media/news-consumption.

- *Oxford Dictionaries*, s.v. "broadsheet" and "tabloid." Accessed 5.3.2018. *https://en.oxforddictionaries.com*.
- Rodgers, Peter, Gem Stapleton, and Peter Chapman. "Visualizing Sets with Linear Diagrams." *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, no. 6 (Dec 14, 2015): 1-39. doi:10.1145/2810012. *http://dl.acm.org/citation.cfm?id=2810012*.
- Rodriguez, Salvador. "U.S. judge says LinkedIn cannot block startup from public profile data." *Reuters.* Published 14.8.2017. Accessed 31.1.2018. *https://www.reuters.com/article/us-microsoft-linkedin-ruling/u-s-judge-says-linkedin-cannot-block-start-up-from-public-profile-data-idUSKCN1AU2BV*.
- Smith, Kevin, M.L.S., J.D. Lecture notes from "Copyright for Educators & Librarians." Duke University, Coursera. Accessed 16.8.2014. *https://www.coursera.org/learn/copyright-for-education*.
- Spence, Robert, and Mark Apperley. "The Encyclopedia of Human-Computer Interaction: Bifocal Display." Interaction Design Foundation. Accessed 21.3.2018. *https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/bifocal-display.*
- Sweney, Mark. "More than half of Britons access news online." *The Guardian*. Published 8.8.2013. Accessed 31.1.2018. *https://www.theguardian.com/media/2013/aug/08/half-britons-access-news-online*.
- Tufte, Edward R. *Beautiful Evidence*. 1. print. ed. Chesire, Conn: Graphics Pr., 2006, 141-155.
- Tufte, Edward R. *Visual Explanations*. 3. print., with rev. ed. Cheshire, Conn: Graphics Pr., 1998, 201-203.
- United States District Court: Northern District of California. *hiQ Labs, Inc v LinkedIn Corporation.* Last modified 14.8.2017, accessed 31.1.2018. *https://www.cand.uscourts.gov/.../C-17-3301-hiQ-v.-LinkedIn-Order-Docket -No.-63.pdf*.
- Wikipedia. "List of newspapers in the United Kingdom by circulation." Last modified 12.3.2018. Accessed 6.11.2017. *https://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_Kingdom_by_circulation*.
- Xin, Li. "Mapping the recent past: Visualization of online news archives." Accessed 6.3.2018. *http://www.thoughtbird.com/portfolio/thesis/*.

10. Tables and Figures

Tables

Figures

11.Appendix

All code and documentation can be found online at https://github.com/lauramatilda/newsScraping.

Sample scraping script

```
import datetime
import urllib
from urllib.request import urlopen, Request
import time
from bs4 import BeautifulSoup
import json
from newspaper import Config, Article
from pymongo import MongoClient
import os

MONGODB_URL = os.environ.get('MONGODB_
URI', 'mongodb://localhost:27017/news')

client = MongoClient(MONGODB_URL)
db = client.get_default_database()
articles = db.articles

USERAGENT = "Mozilla/5.0 (Windows NT 10.0;
Win64; x64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/61.0.3163.100 Safari/537.36"

config = Config()
config.http_success_only = False
config.verbose = True
config.browser_user_agent = USERAGENT

def fetch_page(url):
    url = 'http://www.dailymail.co.uk' + url
    return urlopen(url)

def fetch_article_list(url,d):
  html = fetch_page(url)
  soup = BeautifulSoup(html, 'html.parser')
  html_articles = soup.find('ul', class_='
   archive-articles').find_all('a')
     for html_article in html_articles:
        a = articles.find_one({'url':
  html_article['href']})
  if (('/news/' in html_article['href']) or
  ('/wires/' in html_article['href']) or
  ('/money/' in html_article['href'])):
  if a is None:
     article = {
        'publication': 'daily_mail',
        'method': 'n3k',
        'url': html_article['href'],
        'date': d,
        'title': html_article.get_text()
       }
```

```
    articles.insert_one(article)
    else:
      print('Already indexed',
html_article['href'])
      else:
          print('not relevant arti-
cle', html_article['href'])

def fetch_article_detail(url,article):
    url = 'http://www.dailymail.co.uk' + url
    a = Article(url, config)
    try:
      a.download()
      time.sleep(1)
      a.parse()
    except (KeyboardInterrupt, SystemExit):
      print("!! INTERRUPT !!")
      raise
    except:
      print("!! DOWNLOAD ERROR / PASS !!")
      pass
    else:
      a.nlp()
      article['fetched'] =
      datetime.datetime.utcnow()
      article['pubdate'] = a.publish_date
      print("FETCHED: ",
      article['date'], " ", url)
      article['text'] = a.text
      article['img'] = a.top_image
      article['keywords'] = a.keywords
      article['html'] = a.html
      articles.save(article)

def fetch_detail_loop():
    for article in articles.find():
      if 'fetched' not in article:
          fetch_article_detail(ar-
ticle['url'],article)

def fetch_list_loop():
    datetime_start =
    datetime.datetime(2010, 1, 1)
    offset = 0
    while True:
      d = datetime_start + date
      time.timedelta(offset)
      if d.year >= 2017:
          break
      list_url = '/home/sitemaparchive/
      day_%s.html' % d.strftime('%Y%m%d')
      fetch_article_list(list_url,d)
      offset += 1
      time.sleep(1)

if __name__=='__main__':
    import sys
    if len(sys.argv) < 2:
      print("Usage: python3
      dailymail-n3k.py [list|detail]")
```

```
    sys.exit()
cmd = sys.argv[1]
if cmd == 'list':
    fetch_list_loop()
elif cmd == 'detail':
    fetch_detail_loop()
else:
    print("wrong command")
if cmd == 'list':
    fetch_list_loop()
elif cmd == 'detail':
    fetch_detail_loop()
else:
    print("wrong command")
```


Sample Flask API

```
import json
from datetime import datetime
from dateutil.relative-
delta import relativedelta
from flask import Flask, render_
template, request
from pymongo import MongoClient
from bson.objectid import ObjectId
import os

class JSONEncoder(json.JSONEncoder):
    def default(self, o):
        if isinstance(o, ObjectId):
            return str(o)
        elif isinstance(o, datetime):
            return o.isoformat()
        return json.JSONEn-
coder.default(self, o)

MONGODB_URL = os.environ.get('MONGODB_
URI', 'mongodb://localhost:27017/news')

client = MongoClient(MONGODB_URL)
db = client.get_default_database()
articles = db.articles

app = Flask(__name__)
app.config['DEBUG'] = False
```

```
@app.route("/")
def index():
    article_list = articles.find({
    'publication':'daily_mail',
    'url': {"$regex": "^/news/"}
    })[:50]
    return render_template('arti-
cle_list.html', articles=article_list)

@app.errorhandler(500)
def page_not_found(error):
    return 'This page does not exist', 500

@app.route('/api/<pubName>/<year>/<month>')
def api_query(pubName, year, month):
    start = datetime(int(year), int(month), 1)
    end = start + relativedelta(months=1)
    article_list = articles.find({
    'publication': pubName,
    'text': {"$exists": 1},
    'date': {"$gte" : start, "$lt": end}
    })[:500]
    text_set = []
    for article in article_list:
        text_set.append(article['text'])
    return apiResponse(list(text_set))

def apiResponse(data):
    headers = [["Access-Con-
trol-Allow-Origin", "*"]]
    return (json.dumps(data, cls=-
JSONEncoder), 200, headers)

@app.template_filter()
def nl2br(s):
    return s.replace('\n', '<br>')

@app.template_filter()
def squeeze_breaks(s):
    return s.replace('<br><br>', '<br>')

if __name__ == '__main__':
    port = int(os.environ.get("PORT", 5000))
    app.run(host='0.0.0.0', port=port)
```

Acknowledgements

This work would not have been possible without these generous collaborators, my deepest heartfelt thanks for helping to make this project a reality.

Masood Masoodian - for guidance, direction, and mountains of reading.

Frederik de Bleser & the EMRG - for extensive help setting up the necessary code and infrastructure for this project during a 2017 visit to Antwerp, and for continued patience and guidance in subsequent Helsinki Nodebox workshops.

Martti Niemi - for countless cups of coffee, endless hours of Linux and database troubleshooting, thought-provoking discussions on methods and ethics, and not least of all, moral support and encouragement.

THANK YOU