Popular and on the Rise—But Not Everywhere: COVID-19-Infographics on Twitter

Benedict Witzenberger, Angelina Voggenreiter, and Jürgen Pfeffer

Abstract The coronavirus pandemic has altered many industries around the world. Journalism is one of them. Especially data journalists have gained attention within and outside of their newsrooms. We aim to study the prevalence of journalistic data visualizations before and after COVID-19 in 1.9 million image posts of news organizations on Twitter across six countries using a semi-manual detection approach. We find an increase in the shares of tweets containing infographics. Although this effect is not consistent across countries, we find increases in the prevalence of COVID-19-related content and interactions in infographics throughout all geographies. This study helps to generalize existing qualitative research on a larger, international scale.

Keywords COVID-19 · Data visualization · Data journalism

1 Introduction

COVID-19 served as an accelerator for ongoing changes in journalism: the decline of print and other forms of "traditional" media, the rise of "alternative" news channels, altered skill requirements for journalists, changes in audience and their expectations [19]. Data journalists were central to some of these innovations, as they had the experience and technical means to create data visualizations and exploratory pieces on possible scenarios, which has increased awareness and accessibility to the numbers and fostered engagement of the audience [17]. We define data journalism as the use of data, quantitative analysis, and visualization methods to create journalism [2].

To shed light on the changes that were going on in journalistic data visualization around the world, we analyzed infographics shared by news media before and after COVID-19 hit as a proxy for the prevalence of data journalism. We define information graphics (short: infographics) as a graphical composition of one or more visualizations based on numerical data, images, and text [4].

We aim to answer the following research questions:

RQ1: How has the use of journalistic infographics changed during COVID-19?
RQ2: Which change in the prevalence of journalistic infographics can be found across different countries?
RQ3: How large is the portion of COVID-19 related infographics?
RQ4: How do tweet interactions change in tweets containing infographics compared to other image tweets?

B. Witzenberger (✉) · A. Voggenreiter · J. Pfeffer
School of Social Sciences and Technology, Technical University of Munich, Munich, Germany
e-mail: benedict.witzenberger@tum.de
URL: https://www.hfp.tum.de/

A. Voggenreiter
e-mail: angelina.voggenreiter@tum.de

J. Pfeffer
e-mail: juergen.pfeffer@tum.de

P. Haber et al. (eds.), *Data Science—Analytics and Applications*,
https://doi.org/10.1007/978-3-031-42171-6_7

2 Literature Review

We place our research on the influence of infographics in the news and the impact of COVID-19 on data journalism.

Data journalism during COVID-19 Data journalism is regarded as a form of content or genre innovation in journalism [7], which might provide news companies with an increased reputation or a competitive advantage. Demand for data journalistic training has increased, and data journalists have grown in power and reframed their roles and identity in the newsroom [10]. A positive attitude towards data journalism in the newsrooms correlates with enjoying working with numbers and the belief that competency for data work is satisfactory [3].

The number of published data visualizations during COVID-19 led to criticism about an "information overload" [11], a "bombardment" with visualizations [8] and a very small number of sources, as governmental actors were the main data providers during COVID-19 [14]. Journalists use a small set of authoritative sources, which in turn gain authority by being used in media. To stand out, some media outlets performed their own data collection [5].

The general meaningfulness of leveraging COVID-19-related infographics has been examined in the literature. Visual communication can increase the understandability of highly-scientific topics like the spread of a virus for less educated groups [9], and might lead to higher acceptance for and adherence to protection measures. Research on the effects of infographics during COVID-19 could find positive effects on users' knowledge of mask-wearing techniques, and increased trust compared to text-only guidance, but no substantial effect on COVID-19-related anxiety [6]. Others found a positive effect of infographics on the intention to get a COVID-19 vaccination [20].

Infographics, journalism and COVID-19 Infographics are sometimes described as a method to integrate big data into journalism [21], although this does not necessarily points to the volume of data, which might not really be "big", but aspects like variety (of sources and data types) or veracity of information (in comparison to more traditional ways of reporting). Infographics seem to lead to increased news elaboration and increase more favorable news evaluation [12].

There are three common types of infographics–principle representation (or explanatory visualizations, which explain how things work), cartographic infographics (which show where things are), and statistics charts (which show how many things there are) [22]. We only focus on cartographic and statistic charts here, which are based on some form of numeric data, while principle representations are not necessarily grounded on datasets.

Yet, data journalists have been criticized for COVID-19 charts that might "make the world look more 'fixed' than it really is" [16] with maps not accounting for population densities, models not reporting their underlying assumptions, or exclusion of communities at the margins or the Global South [15].

3 Method

To analyze the diffusion of infographics, we collected data from Twitter, which news media uses mainly as a one-way communication channel to promote reporting [13]. We then implemented a semi-manual approach for infographic detection.

Twitter Collection We collected accounts for the five largest, national, general-audience news media across six different countries by circulation: USA, UK, Germany, France, Italy.[1] To allow some variety of cultural backgrounds while still allowing authors to manually code and understand the content, three English-speaking newspapers from India were included. See Appendix A for accounts.

Tweets were retrieved using `from:USERNAME has:images` on Twitter's API v2 [18]. 2,205,025 tweets for this query were collected for the time range between January 1st, 2018, and July 31st, 2022, in the time between August 15th and September 3rd, 2022. However, contrary to the expected returns, not all of these contained images. In total, we could download 1,911,496 images for analysis, either in JPEG or PNG format. The time range was selected to allow comparable time periods before and after the COVID-19 pandemic started.

Identifying Characteristics of Infographics In line with the definition above, we defined infographics as images containing infographic elements, such as diagrams, maps, or explanatory illustrations, as well as text. To be able to detect infographics

[1] Sources for circulation numbers: Alliance for Audited Media (USA) via pressgazette.co.uk/news/us-newspaper-circulations-2022, ABC (UK): www.abc.org.uk, IVW e. V. (Germany): www.ivw.de, ACPM (France): www.acpm.fr, FIEG (Italy): www.fieg.it, ABC (India): www.auditbureau.org, RNI: rni.nic.in.

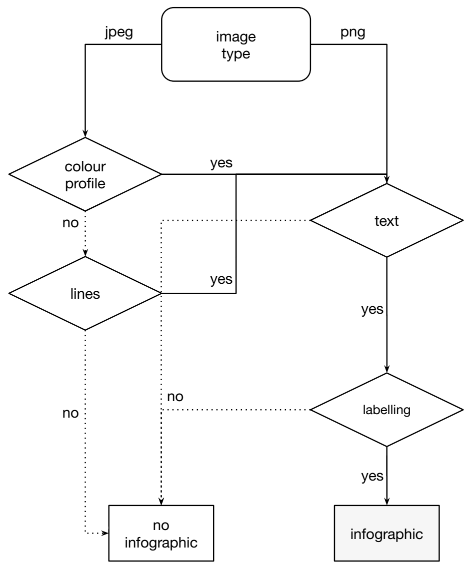

Fig. 1 Infographic detection and labeling process

within our dataset, we first had to define the typical characteristics of an infographic. Therefore, we created a labeled test set consisting of 600 infographics and 1000 non-infographics (as typically more non-infographics than infographics are published by media accounts). The size of the test set allowed us to include a wide variety of images with very different image characteristics, as well as to include non-infographics, which looked very similar to infographics, and vice versa. Using this test set, we identified typical characteristics of infographics and optimized the image characteristic parameters in a way so that **all** infographics of the test set would be extracted (to the price of non-infographics being detected as well):

– **Image type**: If the image was a .png-file, it was likely to be created or edited on a computer and, therefore, likely no common photograph but an image containing graphical elements. For example, 39% of infographics in the test set were PNG-images, while only 14% of non-infographic were PNG-images. We consequently extracted all .png-images [1].
– **Colours**: While photographs typically contain a wide color range, as shadows and light conditions create many different shades of colors in objects, infographics typically consist of a few different graphical elements in a few different colors. It should be noted that transitions between graphical elements can also result in many different color shades, but these colors typically span only a few pixels. We thus extracted all images consisting of few colors spanning a wide area of pixels. In particular, we calculated an RGB histogram of the grayscale image, selected the maximum amount of pixels *pix_max* one color would span, detected the number of colors n spanning at least one-quarter of *pix_max* pixels and extracted the image if $n > 30$. While 98% of infographics in the test set fulfilled this attribute, only 55% of non-infographics did so.
– **Edges**: Infographics mostly contain graphical elements such as text boxes or diagram axes, which can be detected by determining the existence of longer lines in the image. Thus, we extracted all images which included a line spanning at least one-sixth of the minimum of the image length and width. While 94% of infographics in the test set contained such a line, only 64% of non-infographics did so as well.

By extracting all images, which would include at least one of these image characteristics, **all** infographics of the test set could be extracted (resulting in 80% of non-infographics being pulled as well).

From these images, we extracted all images which contained text. We used the *pytesseract* optical character recognition package[2] to detect the existence of text within the image, but as sometimes text in specific fonts was not recognized by the system, around 10% of infographics in the test set were excluded by this step. At the same time, this step was crucial to exclude more non-infographics, as 64% of non-infographics in the test set could be excluded after this step.

Labelling Images After identifying the characteristics of an infographic, we applied these to the dataset of Twitter images by using a semi-automatic approach, as illustrated in Fig. 1: First, we extracted all images, including at least one of the infographic

[2] https://pypi.org/project/pytesseract/.

Table 1 Comparison of predicted labels (with semi-automatic approach) versus human-coded, actual labels in a testset with $n = 2500$

	Actual condition	
Predicted condition	Infographics	Non-infographics
Infographics	26	4
Non-Infographics	8	2462

characteristics described above (image type, colors, edges), which reduced the initial dataset of our Twitter images to 71%. Second, we extracted all images containing text, which decreased the complete dataset size to 16% of the initial dataset. Thirdly, as mentioned above, this process would allow us to detect most infographics at the price of extracting a large number of non-infographics as well. These non-infographics had to be excluded by human inspection. Consequently, we distributed the remaining images to four trained human annotators, who manually excluded all images not being an infographic. The annotators were instructed to focus on cartographic or statistical charts, which needed to be grounded on numeric data and were not solely a text containing a single number, but a visual representation of data. The step of manual inspection reduced the dataset size to 1% of the initial size.

Finally, we evaluated this semi-automatic approach by creating a random subset of 2,500 images, labeling these images with the semi-automatic approach (to infer the predicted labels), manually inspecting these images (to infer the true labels), and comparing the labels. Our approach showed a sensitivity of 0.765 (26 out of 34), a specificity of 0.998 (2462 out of 2466), an accuracy of 0.995, and an F1-score of 0.813 (see Table 1).

4 Results

Out of the 1,911,496 images we analyzed, we found 25,813 infographics using the semi-automatic approach described above.

An increase in infographics—but not everywhere The share of infographics within all shared media images increased significantly after the COVID-19 pandemic hit. We defined 'after COVID-19' by tweets published after March 1st, 2020—when the pandemic as a journalistic topic had spread worldwide. We found an increase of 42% between pre- and post-pandemic infographic proportions from 1.2% ($n = 10{,}652$) to 1.6% ($n = 15{,}161$) ($\chi^2 = 742.98$, $p < 0.01$, $df = 1$). Still, their share within all images was just around 1.6%, with most images (98.4%) remaining non-infographics. While this seemed to be a clear direction, we found differences when splitting the data into the observed countries.

Not all countries had similar increases in infographics. We found that media in the US had the largest absolute increase of infographics with 2.1 percentage points from 5.7% before the pandemic to 7.8% after. The highest proportional increase was found in India, which increased the share by 130% from 1% to 2.3%. In the UK, we found similar relative growths of around 125%.

In contrast, the share of infographics remained constant in German media at 0.7%. In Italy and France, we found fewer infographics after COVID-19: Italy decreased by around 10% to 1.9%, and in France, the share dropped from 2.8% before COVID-19 to 1.9% after, reducing by around 32%.

COVID-19 is a prominent topic in tweets To further understand the content of the infographics, we analyzed the 50 most-used hashtags for each country. These were manually coded into six categories: COVID-19, politics, ukraine, elections, sports, and others. Ambiguous terms were added to the most distinguishing category ("Biden" to 'elections', "putin" to 'ukraine'). As COVID-19 brought up a set of new, distinctive words, its category seemed very unequivocal.

COVID-19 was found in between 8.1 and 23% of all tweets in our sample after the start of the pandemic. This also holds if only regarding infographics tweets, where COVID-19-related infographics made up 27.8% ($n = 4,220$) of all 15,161 detected infographics. While everywhere the share of tweets about COVID compared to other topics was higher for non-infographics than for infographics, this difference was very small in France (11.1% infographics, 13% non-infographics COVID-19-related). However, 63% of tweets could not be attributed.

When only regarding infographics after COVID-19 hit, we found that nearly half of infographics in German media were COVID-related, 33.2% in Italy, 30.8% in India, and 26.2% in France (see Fig. 2). The UK and especially the US had smaller shares for COVID-19.

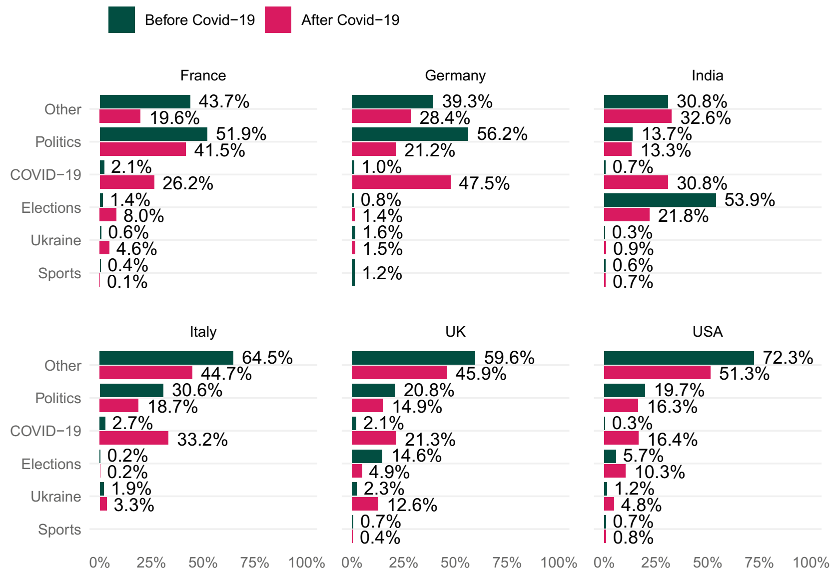

Fig. 2 Share of infographic tweets before and after COVID by clustered topics

Infographic tweets receive higher interactions From a social media perspective, infographics are a valuable tool for media organizations that seek to promote their content via Twitter. We found that mean counts for likes (37.12 versus 66.14), retweets (10.05 versus 28.89), quotes (3.01 versus 6.65), and replies (6.05 versus 7.11) were significantly higher for tweets containing an infographic ($p < 0.001$).

The audience's interest was also visible when comparing infographics before and after COVID-19. We found that after COVID, mean likes (31 versus 90), quotes (5.2 versus 7.6), and replies (4.8 versus 8.7) counts are significantly greater than before ($p < 0.001$). Retweets showed a non-significant increase from 21.2 to 34.3. While we cannot account for changes in followers, as data collection took place at a fixed point in time, we could see that the increases in likes, retweets, and quotes were much higher for infographics than for non-infographics.

5 Discussion

COVID-19 had a solid effect on the prevalence of journalistic infographics on Twitter. We have found an increase in the use of infographics on media's Twitter pages after the start of the COVID-19 pandemic overall, addressing **RQ1**. This increased output of infographics is in line with qualitative literature [8, 11] that labeled these visualizations an "infodemic". Compared to the total number of tweets, however, infographics still only make up a small portion.

For **RQ2**, we found differences between countries. While there were substantial increases for US, Indian, and UK media, stagnation occurred in German media, and we surprisingly found slight declines in France and Italy. We can rule out some possible explanations for this: First, there might be a general difference in the reception of COVID-19 across the studied countries. However, we found much COVID-19-related content when analyzing the infographic tweets' texts across all countries–which also addresses **RQ3**. This leads us to believe that there was a consensus among journalistic infographic designers to produce COVID-19-related content in all observed countries. Second, some media might have different strategies for promoting their content on Twitter. While we cannot control for this from the outside perspective we have taken, we can show for **RQ4** that image tweets containing an infographic receive higher interactions. From a media distribution standpoint, it is rational to use these graphics on Twitter.

Nonetheless third, some media might have different approaches to posting data on social media. While manually labeling the automatically detected infographics, we found "text boards" in many instances. A computational, infographic-like image that contains numbers in textual format, mostly combined with images. These can be regarded as having an infographical appearance and could serve as a substitute for creating charts, which leads to higher requirements for data collection and analysis. From a definitional point of view, we decided not to include these images, as they do not contain charts but only display single numbers. As we only attributed binary labels to the infographics, we cannot control whether this has had a huge influence on the analysis.

6 Limitations

To explain precisely what led to the stagnation in Germany and the decline in Italy and France would require more insights into the newsrooms to rule out editorial decisions which are not visible from the outside. Some qualitative work has already been accomplished around this [5], restricted in scope and generalizability, however, by the efforts that a qualitative study requires.

This study is limited by several factors: Our approach left us relying on Twitter to detect images correctly. As we collected all data at one point in time, tweets that had been deleted could not be used for this study. The focus on six countries might have included strong influences of western-democratic media business that might not be applicable elsewhere. Although Indian media also followed the trend, restricting it to English-speaking media might have influenced the outcome, as others have discovered differences between western democracies and the Global South [14, 15].

Further research might focus on a larger variety of non-western countries to enhance understanding of possible differences in media cultures around the publication of infographics. It might also be beneficial to develop quantitative methods to detect publication differences within certain media markets, which is a field that is mostly covered by qualitative work and is hard to generalize and transfer to other populations.

In addition, our semi-automatic approach was restricted by the quality of text detection. As described in Sect. 3, the existence of text within an image was a critical, required factor in differentiating between infographics and non-infographics, but at the same time, text detection failed in around 10% of infographics. Future research could use more elaborated text detection techniques to also take small, hardly readable, and non-standard texts (e.g., text in the form of word art) into account. This also limits our results which did not include infographics without text, which, however, we only expect to appear in very few cases, as data visualizations usually need some form of textual integration.

7 Conclusion

COVID-19 has influenced innovation in journalism in a lot of ways. We presented a quantitative study on the use of infographics on Twitter before and after COVID-19, which confirms earlier qualitative research. We saw an increase in infographics in our sample of image posts by the largest newsrooms in three of the six researched countries. However, in some countries, they declined. Nonetheless, we found an increase in COVID-19-related content, which is high across all geographies studied for image tweets in general, and infographic tweets in particular. Interactions for infographic tweets are higher than for images. This remains a topic for further research with deeper insights into newsroom practices. News organizations could adapt to COVID-19 in various ways. Increased use of infographics is just one of these developments–and might further influence reporting with potential consequences on the perception of journalism.

Acknowledgements The authors would like to thank Sophie Brandt, Linus Mosch, and Sarah Talbi for their help with labeling the visualizations.

A Media Accounts

See Table 2.

Table 2 Selected accounts for analysis

Country	Accounts	Followers
France	L'Humanite (@humanite_fr)	396 k
	Le Figaro (@le_figaro)	3.6 m
	Le Monde (@lemondefr)	10.5 m
	Le Parisien (@le_Parisien)	3.2 m
	Liberation (@libe)	3.4 m
Germany	Bild (@BILD)	1.9 m
	Die Welt (@welt)	1.8 m
	Frankfurter Allgemeine (@faznet)	803 k
	Handelsblatt (@handelsblatt)	379 k
	Sueddeutsche Zeitung (@SZ)	1.8 m
India	Hindustan Times (@httweets)	8.6 m
	The Hindu (@the_hindu)	7.9 m
	Times of India (@timesofindia)	14.6 m
Italy	Corriere della Sera (@Corriere)	2.7 m
	Il Resto del Carlino (@qn_carlino)	55 k
	Il Sore 24 Ore (@sole24ore)	1.8 m
	La Reppublica (@repubblica)	3.5 m
	La Stampa (@lastampa)	1.3 m
UK	Daily Mail (@mailonline)	2.8 m
	Daily Mirror (@dailymirror)	1.3 m
	The Daily Telegraph (@telegraph)	3.3 m
	The Sun (@thesun)	2 m
	The Times (@thetimes)	1.7 m
USA	Los Angeles Times (@latimes)	4 m
	New York Times (@nytimes)	54.9 m
	USA Today (@USATODAY)	4.9 m
	Wall Street Journal (@wsj)	20.4 m
	Washington Post (@washingtonpost)	20 m

References

1. Adler, M., Boutell, T., Brunschen, C., Costello, A.M., Crocker, L.D., Dilger, A., Fromme, O., Gailly, J.l., Herborth, C., Jakulin, A., Kettler, N., Lane, T., Lehmann, A., Lilley, C., Martindale, D., Mortensen, O., Pickens, K.S., Poole, R.P., Randers-Pehrson, G., Roelofs, G., van Schaik, W., Schalnat, G., Schmidt, P., Wegner, T., Wohl, J.: PNG (Portable Network Graphics) Specification. Version 1.0. W3C (1996) https://www.w3.org/TR/REC-png-961001
2. Anderton-Yang, D., Kayser-Bril, N., Howard, A., Teixeira, C.V., Slobin, S., Vermanen, J.: Why is data journalism important? In: Gray, J., Bounegru, L., Chambers, L., European Journalism Centre, Open Knowledge Foundation (eds.) The Data Journalism Handbook 1. European Journalism Centre (2012) https://datajournalism.com/read/handbook/one/introduction/why-is-data-journalism-important
3. Appelgren, E., Nygren, G.: Data journalism in Sweden. Digit. J. **2**(3), 394–405 (2014). https://doi.org/10.1080/21670811.2014.884344
4. Cairo, A.: The Functional Art. New Riders Publishing (2012)
5. Desai, A., Nouvellet, P., Bhatia, S., Cori, A., Lassmann, B.: Data journalism and the COVID-19 pandemic: opportunities and challenges. Lancet Digit. Health **3**(10), 619–621 (2021). https://doi.org/10.1016/S2589-7500(21)00178-3
6. Egan, M., Acharya, A., Sounderajah, V., Xu, Y., Mottershaw, A., Phillips, R., Ashrafian, H., Darzi, A.: Evaluating the effect of infographics on public recall, sentiment and willingness to use face masks during the COVID-19 pandemic: a randomised internet-based questionnaire study. BMC Pub. Health **21**(1), 367 (2021). https://doi.org/10.1186/s12889-021-10356-0
7. García-Avilés, J.A.: Reinventing television news: Innovative formats in a social media environment. In: Studies in Big Data, pp. 143–155. Springer International Publishing (2020). https://doi.org/10.1007/978-3-030-36315-4_11
8. García-Avilés, J.A., Arias-Robles, F., de Lara-González, A., Carvajal, M., Valero-Pastor, J.M., Mondéjar, D.: How COVID-19 is revamping journalism: newsroom practices and innovations in a crisis context. J. Pract. 1–19 (2022). https://doi.org/10.1080/17512786.2022.2139744
9. Hamaguchi, R., Nematollahi, S., Minter, D.J.: Picture of a pandemic: visual aids in the COVID-19 crisis. J. Pub. Health **42**(3), 483–485 (2020). https://doi.org/10.1093/pubmed/fdaa080
10. Hermida, A., Young, M.L.: Data Journalism and the Regeneration of News (2019). https://www.routledge.com/Data-Journalism-and-the-Regeneration-of-News/Hermida-Young/p/book/9781138058934

11. Krawczyk, K., Chelkowski, T., Laydon, D.J., Mishra, S., Xifara, D., Gibert, B., Flaxman, S., Mellan, T., Schwämmle, V., Röttger, R., Hadsund, J.T., Bhatt, S.: Quantifying online news media coverage of the COVID-19 pandemic: Text mining study and resource. J. Med. Internet Res. **23**(6), e28253 (2021). https://doi.org/10.2196/28253
12. Lee, E.J., Kim, Y.W.: Effects of infographics on news elaboration, acquisition, and evaluation: Prior knowledge and issue involvement as moderators. New Media Soc. **18**(8), 1579–1598 (2016). https://doi.org/10.1177/1461444814567982
13. Malik, M.M., Pfeffer, J.: A macroscopic analysis of news content in twitter. Digit. J. **4**(8), 955–979 (2016). https://doi.org/10.1080/21670811.2015.1133249
14. Mellado, C., Georgiou, M., Nah, S.: Advancing journalism and communication research: new concepts, theories, and pathways. J. Mass Commun. Quart. **97** (2020). https://doi.org/10.1177/1077699020917204
15. Milan, S., Treré, E.: The rise of the data poor: The COVID-19 pandemic seen from the margins. Soc. Media + Soc. **6**(3) (2020). https://doi.org/10.1177/2056305120948233
16. Northwestern Buffett Institute for Global Affairs: Visualizing a World of COVID-19 Uncertainty (2020) https://buffett.northwestern.edu/news/2020/visualizing-a-world-of-covid-19-uncertainty.html
17. Pentzold, C., Fechner, D.J., Zuber, C.: "Flatten the curve": Data-driven projections and the journalistic brokering of knowledge during the COVID-19 crisis. Digit. J. **9**(9), 1367–1390 (2021). https://doi.org/10.1080/21670811.2021.1950018
18. Pfeffer, J., Mooseder, A., Lasser, J., Hammer, L., Stritzel, O., Garcia, D.: This sample seems to be good enough! assessing coverage and temporal reliability of Twitter's Academic API. In: Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM-2023) Forthcoming (2023)
19. Quandt, T., Wahl-Jorgensen, K.: The Coronavirus pandemic and the transformation of (digital) journalism. Digit. J. **10**(6), 923–929 (2022). https://doi.org/10.1080/21670811.2022.2090018
20. Riggs, E.E., Shulman, H.C., Lopez, R.: Using infographics to reduce the negative effects of jargon on intentions to vaccinate against COVID-19. Pub. Understand. Sci. **31**(6), 751–765 (2022). https://doi.org/10.1177/09636625221077385
21. Smit, G., Haan, Y.D., Buijs, L.: Working with or next to each other? boundary crossing in the field of information visualisation. J. Media Innovat. **1**(2), 36–51 (2014). https://doi.org/10.5617/jmi.v1i2.875
22. Zwinger, S., Langer, J., Zeiller, M.: Acceptance and usability of interactive infographics in online newspapers. In: 2017 21st International Conference Information Visualisation (IV). IEEE (2017). https://doi.org/10.1109/iv.2017.65