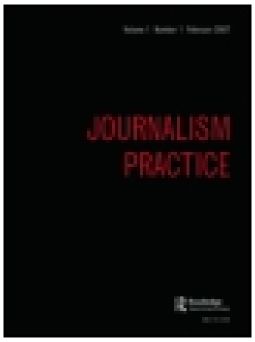

JOURNALISM
PRACTICE

Journalism Practice

Routledge
Taylor & Francis Group

ISSN: 1751-2786 (Print) 1751-2794 (Online) Journal homepage: http://www.tandfonline.com/loi/rjop20

Classifying Data Journalism

Florian Stalph

To cite this article: Florian Stalph (2017): Classifying Data Journalism, Journalism Practice, DOI: 10.1080/17512786.2017.1386583

To link to this article: http://dx.doi.org/10.1080/17512786.2017.1386583

Published online: 19 Oct 2017.

CLASSIFYING DATA JOURNALISM
A content analysis of daily data-driven stories

Florian Stalph

The review of theoretical and empirical studies in data journalism has uncovered different conceptualisations of data journalistic artefacts. This quantitative content analysis of data-driven stories published by European quality news websites Zeit Online, Spiegel Online, The Guardian *and* Neue Zürcher Zeitung *aims to outline universal characteristics of daily data-driven stories and to compare these findings with previous analyses of data stories and acclaimed data journalism projects. Results suggest that daily data journalism stories generally feature two visualisations that are likely to be bar charts. The majority of these visualisations are not interactive whereas maps turn out to be the most interactive type of visualisation. Data journalists rely predominantly on pre-processed data drawn from domestic governmental bodies. For the most part, data-driven stories are reports on political topics paralleling traditional news reporting. The sparsity of collaborative efforts and investigative approaches distinguishes daily data journalism from previous analyses of eclectic and elaborate data-driven projects.*

KEYWORDS computational journalism; data-driven journalism; data journalism; data visualisation; digital journalism; Europe

Introduction

Recently, the acclaimed international collaboration based on an extensive analysis of an unprecedented data leak that led to the Panama Papers has been appraised as a prime example of data journalism—and rightfully so. At the same time, stories like this one do not represent the current state of data journalism. As a journalistic discipline, it is still developing just as it still is a rather unstudied field in journalism research. Aside from national case studies and epistemological insights, data journalistic artefacts remain largely unexplored. Two quantitative content analyses of submissions to data journalism awards (Reimer and Loosen 2017; Young, Hermida, and Fulda 2017) construe characteristics of best-practice data journalism. While these studies show what top-notch data journalism looks like, they disregard “daily, quick turnaround, generally visualised, brief forms of data journalism” (Borges-Rey 2016, 6) or “*general data journalism*” (Uskali and Kuutti 2015, 85). The present content analysis sets out to investigate these forms of data-driven articles published via websites of European legacy media (*Zeit Online*, *Der Spiegel*, *The Guardian*, *Neue Zürcher Zeitung*) in order to draw comparisons between acclaimed and everyday articles with the objective of sharpening the image of applied data journalism.

Literature Review

Research related to data journalism can be distinguished between national and institutional case studies and studies on the epistemology of data journalism. Case studies in

Journalism Practice, 2017
https://doi.org/10.1080/17512786.2017.1386583

Routledge
Taylor & Francis Group

the United States (Parasie and Dagiral 2013; Fink and Anderson 2015; Parasie 2015), Canada (Tabary, Provost, and Trottier 2016), the United Kingdom (Borges-Rey 2016), Norway (Karlsen and Stavelin 2014), Sweden (Appelgren and Nygren 2014), Belgium (De Maeyer et al. 2015) and Germany (Weinacht and Spiller 2014) comprehensively focus on the integration of data journalism into newsrooms and accompanying challenges and implications for news outlets and journalists. Generally, there is no unanimous definition of data journalism across institutions on staff and managerial levels (Appelgren and Nygren 2014; De Maeyer et al. 2015; Fink and Anderson 2015; Karlsen and Stavelin 2015). These structural and conceptual uncertainties lead to diverse role perceptions as well as data journalistic products (Weinacht and Spiller 2014; De Maeyer et al. 2015; Borges-Rey 2016), thereby impeding the access to data journalism as a research object. The lack of a theoretically and practically acknowledged concept results in various interpretations of requirements considering editorial environments and skill-sets. Therefore, data journalists are facing time pressures, a general lack of resources as well as insecurities regarding computational or analytical skills (Karlsen and Stavelin 2014; Appelgren and Nygren 2014; De Maeyer et al. 2015; Fink and Anderson 2015; Tabary, Provost, and Trottier 2016). Besides these practical insights into the application of data journalistic methodologies, epistemological examinations discuss the scope of data journalism, its genesis and inherent influences on existing journalistic systems (Parasie and Dagiral 2013; Gynnild 2014; Coddington 2015; Parasie 2015; Borges-Rey 2017). Despite data artefacts being the constituent factor of data journalism, the significance of investigating data, collecting data, scrutinising data providers (Parasie and Dagiral 2013; Tabary, Provost, and Trottier 2016) and the influence of data on the editing process (Borges-Rey 2016) remains undetermined. Furthermore, these studies discuss the effects of computational approaches and technological progress on traditional journalism, considering data journalism as a manifestation incorporating these constituents (Karlsen and Stavelin 2014; Parasie and Dagiral 2013; Gynnild 2014; Parasie 2015; Borges-Rey 2017). Moreover, data journalism as a new form of interdisciplinary journalism needs to become aware of its role in journalistic systems: rooting in digital advocacy, data journalism has to determine its function as an intermediary, linking socio-political responsibilities of journalism to digital spheres (Parasie and Dagiral 2013; Gynnild 2014; Parasie 2015). Hitherto, quantifiable research results related to data journalism are scarce due to reasons pointed out above.

Besides phenomenological exploratory case studies and epistemological discussions, data journalistic artefacts, i.e. articles edited by data journalists, remain widely unstudied. Within their combined study, Parasie and Dagiral (2013) analysed data-driven stories from the printed edition of the *Chicago Tribune* ($N = 69$). They found that 60.8 per cent of all stories rely on data published by public officials. Underlying data-sets are used as provided through governmental releases without in-house reporters creating databases. Based on these results, the authors suggest reducing "the dependence on government agendas" (13). Furthermore, they could identify a strong focus on education topics (46.4 per cent) and articles based on demographic data (20.3 per cent). Last, Parasie and Dagiral determined that solely 10.1 per cent of the stories do not feature charts, tables or maps (6). Besides their case study, Tabary, Provost, and Trottier (2016) also examined 178 data journalism projects from Quebec. Their results imply a high "dependency on pre-processed public data" (75): only 2 per cent is original data whereas almost half of all data-sets are institutional and 16 per cent a mix of institutional and other data sources. Regarding visualisations, 58 per cent of all graphics are maps while merely half of these offer interactive added value through quantitative information (78).

So far, there are four dedicated content analyses that examine the structure, visualisations, sources and methodologies of data journalism stories. Similar to this study, Knight's (2015, 69) content analysis of data-driven stories published by mainstream news media in the United Kingdom ($N = 106$) aimed at comparing "claims made by data journalism evangelists against the reality". The results show that quality news outlets such as *The Guardian* produce more complex and quality data-driven stories than tabloid newspapers. Infographics were found to be the most popular means to represent data, followed by static maps, graphs or charts, and number pull quotes. In accordance with other content analyses, institutional sources, particularly government agencies, account for the majority of used data sources whereas press releases by research institutes are often referred to by social issue and health stories. Furthermore, investigative approaches were hardly observable (Knight 2015, 62–69).

Tandoc and Oh (2017, 1004–1005) analysed "new stories that utilized big data" published via *The Guardian*'s *Datablog* ($N = 260$) and found that these stories heavily rely on governmental sources (29.7 per cent), followed by data generated by the news organisation itself (18.5 per cent). Besides examining news values and topics of these stories, they found that tables, photographs and static infographics are the most employed visualisations (1008–1009).

Focusing on Canadian entries ($N = 26$) of three data journalism awards,[1] Young, Hermida, and Fulda (2017) concluded that most of the time one or two journalists edit a story, and more than half of the stories could be described as investigative with strong geographical, mostly local, references. Considering visualisations, dynamic maps were identified as the most used technique featuring some interactive components. Half of all visualisations are based on public records while overall more than two-thirds of the used data-sets are not easily accessible by the reader. The study concluded that there appears to be "a lack of clear standards regarding what is considered as excellence in data journalism awards submissions, the degree of interactivity and how the latter is being implemented" (12–13).

Similarly, Reimer and Loosen (2017) examined data stories submitted to the Global Editors Network Data Journalism Award ($N = 179$) between 2013 and 2015. The study shows that data-driven articles have a strong focus on "political, societal and economic issues" (20), while education, culture and sports are relatively underrepresented. In accordance with the Canadian analysis, maps are a common visualisation type, only topped by static charts. Furthermore, Reimer and Loosen pointed out that interactivity is limited to rather unsophisticated features (20). More than two-thirds of the indicated data sources are provided by official institutions—primarily geo-data, financial data and sensor data, followed by socio-demographic data. Interestingly, almost 20 per cent of all data-sets had been collected by media organisations themselves (10–12). Both content analyses provide comprehensive insights into data journalistic articles that have been awarded by or submitted to data journalism award committees—both studies clearly indicate that their intention is to outline the characteristics of alleged best-practice data journalism knowing that "they are likely not to represent 'everyday' data journalism" (22). Also Borges-Rey (2016, 9) concludes that different forms of data journalism are observable: concise data-driven stories done on a daily basis, extensive investigative projects and gamified variants. In a like manner, Uskali and Kuutti (2015, 85) distinguish between "investigative data journalism" and "general data journalism". The present study follows up directly on this debate by providing an extensive analysis of so-called general or daily data journalism. Accordingly, this study examines a sample of randomly selected data journalistic articles published on designated landing pages of European legacy media websites. I will discuss the samples of this present study and of previous content analyses in the following methodology section and contrast the results in a closing comparative discussion below. Based on these premises, the following main research questions are raised:

> **RQ1:** What are formal characteristics of daily data-driven stories published on European legacy media websites?
>
> **RQ2:** In what manner are visualisations employed as a graphical representation of data?
>
> **RQ3:** To what extent do data sources shape stories?
>
> **RQ4:** What are forms and contents of data-driven stories?

The analytical framework comprises four major dimensions that let us examine data-driven articles on two levels. Each dimension focuses on different aspects of data-driven articles to satisfy desiderata of definitional attempts and current research results.

I differentiate between two levels, one examining data journalistic characteristics, i.e. data visualisations and accompanying data sources. On a separate text level, the forms and contents of data-driven stories are examined. These two levels compose the following four dimensions:

1. Formal characteristics.
2. Data visualisations.
3. Data sources.
4. Form and content.

In order to examine data visualisations, data sources, and forms and contents, I operationalised relevant variables and designed empirical categories for the different dimensions of this content analysis. Table 1 provides an overview of the dimensions and their accompanying variables.

Formal characteristics serve to examine physical appearances and systemic attributes of data-driven articles. As regards content, I analyse data visualisations concerning the proportion of text visualisations, visualisation types and interactivity levels of each visualisation according to Schulmeister's (2003) taxonomy of multimedia component interactivity. He developed an interactivity scale to classify didactic components such as "images, diagrams, animations sequences, video clips, audio samples, or tables, formulas, *JavaApplets*, and *Flash* programs" (64) in multimedia learning systems or learning websites. In the present context, I consider data visualisations as didactic elements as readers can derive information and knowledge from data-driven graphics. Schulmeister proposes six levels of interactivity (65–71). At the first level, the least interactive, objects do not provide any interactive features. Level II objects can also be merely viewed, there are, however, multiple pre-defined variations available. Level III allows the user to manipulate the visualisation by scaling or rotating it, choosing different perspectives, or navigating within the graphic. At the fourth level, visualisations are not predefined; the reader can input a query and generate a graphical component. At the fifth

TABLE 1
Analytical framework

Dimension	Variables
Formal characteristics	Date of publication
	Medium
	Length in words
	Number of authors
	Topic
Data visualisations	Ratio text/visualisation
	Number of visualisations
	Visualisation types
	Level of interactivity
Data sources	Provision of data
	Number of sources of each visualisation
	Data provider
	Country of origin
	Accessibility of data
Form and content	Story format
	Subject matter
	Foreign news
	Juxtaposition

level, the user can build a viewing object from scratch using a comprehensive toolset. Level VI adds feedback to the previous level that is computed by an intelligent program based on expanding symbolic objects into meaningful objects. This level is principally subject to mathematical or geometrical programs.

Up to eight visualisations of each article have been taken into account. If an article contains more than eight graphical representations of data, they are counted but not analysed regarding visualisation type and level of interactivity. Considering data sources, I examine whether or not data-sets are provided to readers, count the number of sources the authors used to create a visualisation, and identify the institution or organisation that carries the data and the national origin of the data source.

Last, I analyse forms and content of a story on a separate level. I examine the format and subject matter of each article, whether there are references to foreign news or if the stories exclusively cover domestic news. Additionally, I explore whether there is a perceivable contrasting juxtaposition in terms of an object of investigation that is being scrutinised, hinting at investigative approaches.

Sampling

A sample consisting of 244 data-driven articles published via websites of four European legacy media outlets was compiled. I selected outlets that publish data stories in the German or English languages. As there is still no universally accepted definition of data journalism that would specify the features of a data journalistic product, sampling data-driven articles proves difficult. Most websites do not have specific data journalism sections or even departments, adding to the academic discussion of data journalism being a genre, methodology or storytelling format. As implicated by Weinacht and Spiller (2014), data journalism is employed across all sections; at the same time, there is no unanimous definition of what a data-driven story looks like, thus providing no indicator in this respect. By compiling

TABLE 2
Sample

Website	*Zeit Online*	*Spiegel Online*	*The Guardian*	*Neue Zürcher Zeitung*
Section	*Datenjournalismus*	*Datenlese*	*Datablog, Data Visualisations, Data Journalism*	*NZZ Data*
Sample	61	61	61	61

units of analysis that offer allegedly typical characteristics such as visualisations, indicators of data sources or numerical data in texts, I would disregard stories that do not feature these superficial traits but are based on data-driven investigations nonetheless—these systematic and ultimately epistemological uncertainties have previously been discussed by Tabary, Provost, and Trottier (2016, 73) in a similar manner. Taking this into consideration, in order to avoid a bias based on the form of data-driven pieces that very much vary in their appearance—e.g. data-driven articles may not have visualisations at all—I did not compile the data-set by selecting articles based on the assumption that they would represent certain data characteristics. The determining factor of this present sample is that all data-driven stories have been published on designated data journalism landing pages of the outlets' websites that are explicitly labelled as such. *Zeit Online*, website of the German weekly newspaper *Die Zeit*, publishes data-driven stories via a landing page entitled *Datenjournalismus*. The German weekly print magazine's Web presence *Spiegel Online* accumulates respective articles under the label *Datenlese*. *The Guardian* deposits their data-driven stories in the columns *Datablog*, *Data Visualisations* and *Data Journalism*. The website of the Swiss daily newspaper *Neue Zürcher Zeitung* publishes relevant articles on their landing page *NZZ Data*.[2]

The population consists of all articles published via the above-mentioned landing pages between September 2013 and September 2015 and that were retrievable and accessible at the due date of data collection (18 January 2016). The final sample consists of 61 randomly selected sampling units per news site (see Table 2).

I reckon that this sample will allow us to gain comprehensive insights into the structure and systematic of daily data-driven stories. Deliberately selecting a random sample suggests more generalisable implications than previous content analyses of data journalism artefacts. At the same time and by following this approach, I cannot eliminate the possibility that units of this sample could also be stories submitted to data journalism awards or stories that meet characteristics of those regarding complexity or employed means. As the present sample is drawn from established legacy media that maintain well-equipped data teams which have been nominated for data journalism awards, the two opposing categories of data-driven stories submitted to awards and daily data stories are ideal-typical. The samples of previous studies as well as of this analysis should be considered approximations to either of these categories.

Results

A First Approximation: Length, Authors, Topics, Visualisations

Based on the results, the following formal characteristics of data-driven articles can be outlined. The data show that the stories are of 575 words in length on average

TABLE 3
Topics of data journalistic stories

Topic	*N*	% of total
Politics	96	39.3
Society	38	15.6
Business	29	11.9
Culture	10	4.1
Sports	19	7.8
Local	19	7.8
Other	33	13.5
Total	244	100.0

(mean = 575.05); 90 per cent of the sampled stories are shorter than 1049 words. However, there is a standard deviation of 705 words as there are bigger data-driven projects in the sample, one with an extreme value of 7270 words. Without regard for this extreme value, there is a slight correction with a mean of 548 words in length. More than half of all stories (59 per cent) are authored by one journalist and one-quarter by a team of two journalists; 6.1 per cent are edited by three or more authors, making larger teams even more uncommon. These results suggest that bigger data-driven projects edited by large, collaborating teams are exceptional cases.

As Table 3 shows, the majority of data-driven stories deal with politics, followed by society and business topics. These findings show that data journalists apply their practices across several sections but data-driven stories predominantly cover politics.

Regarding the style of the articles, 76.2 per cent of the sampled data-driven stories can be described as traditional news reports ($N = 186$), compared to 15.2 per cent ($N = 37$) being features or explanatory stories providing detailed background. In terms of the overall composition, 75.4 per cent ($N = 184$) of 244 stories include both text and visualisation, compared to 12.7 per cent ($N = 31$) being visualisation only and 11.9 per cent ($N = 29$) text only. This shows that data visualisations are a distinctive and very central feature of data-driven stories. In exceptional cases, however, data are not necessarily constituent in the form of a visual representation as they can also be presented in textual form.

Figure 1 shows a detailed distribution of numbers of visualisations grouped by news outlets. After eliminating five extreme values, a mean value of 1.97 (SD = 1.69) visualisations per article is determined.

The first implications I can draw from this initial analysis is that three-quarters of data-driven stories include text of 575 words in length, two visualisations and are edited by one author. With regard to content and structure, the articles mostly appear as traditional reports and cover politics.

Visualisations as Storytelling Elements

Distributed over 244 articles, the sample contains 533 visualisations. I examine the first eight visualisations of each article, effectively reducing the number of analysed visualisations to 510. Bar charts are by far the most used visualisation types in the sample. Almost one-third of all stories rely on a bar chart, followed by maps and line charts. According to

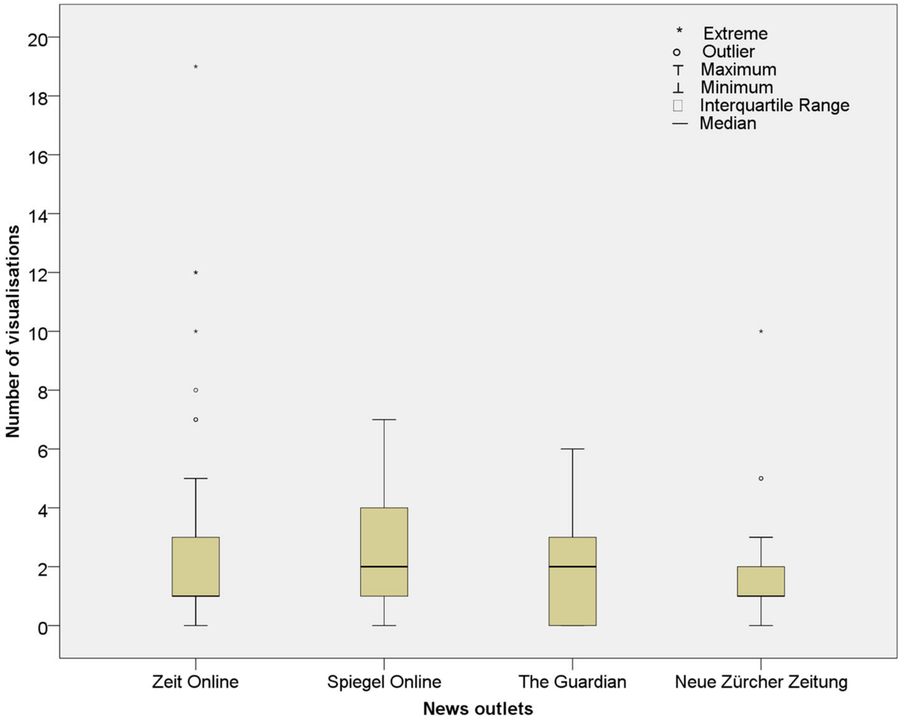

FIGURE 1
Number of visualisations grouped by news outlets

Roam's (2009) showing framework, charts in general (including line charts, pie charts, area charts and radar charts) serve the purpose of showing numerical values by displaying quantities in certain forms that facilitate comparisons. Also, Kirk (2016, 31) refers to bar charts as "the *old reliables* of the field—always useful, always being used, always there when you need them". Regarding their function, he categorises bar charts as the most basic chart types providing "[c]ategorical comparisons" (159). The data confirm the claim of bar charts being the most used visualisation type (Table 4).

Three-quarters of the sampled stories build upon both text and visualisations. Nevertheless, as 11 per cent of the articles do not feature any kind of visualisation, I can support Weinacht and Spiller's (2014, 418) claim that data-driven stories do not necessarily have to implement graphical representations of data. Those pieces coded as "visualisation only" oftentimes do have textual information, however, as overlays, legends or explanatory text fragments within the visualisation in contrast to clearly offset and stand-alone text segments (Table 5).

Visualisations regarding their interactivity are analysed according to Schulmeister's (2003) taxonomy. I categorise almost half of all charts (see Table 6) as level I on the interactivity scale, meaning that the charts "only serve for illustration or information" and the "contents remain constant" (65); 47.6 per cent of all level I charts are bar charts and line charts, making them the least interactive out of the three most used data visualisations. These data representations are not manipulable in any way, thus, these visualisations do not offer interactive features at all. In summary, bar charts appear to be the most used

TABLE 4
Visualisation types

Visualisation type	*N*	% of total	Cumulative contribution
Bar chart	167	32.7	32.7
Map	83	16.3	49.0
Line chart	59	11.6	60.6
Pie chart	41	8.0	68.6
Table	33	6.5	75.1
Infographic	21	4.1	79.2
Pictogram	14	2.7	81.9
Area chart	12	2.4	84.3
Bubble chart	6	1.2	85.5
Scatter plot	5	1.0	86.5
Radar chart	4	0.8	87.3
Timeline	4	0.8	88.1
Flow chart	3	0.6	88.6
Dashboard	42	8.2	96.9
Other	16	3.1	100.0
Total	510	100.0	

TABLE 5
Level of interactivity

Level of interactivity	*N*	% of total	Cumulative contribution
Level I	248	48.6	48.6
Level II	191	37.5	86.1
Level III	51	10.0	96.1
Level IV	14	2.7	98.8
Level V	6	1.2	100.0
Level VI	0	0.0	
Total	510	100.0	

TABLE 6
Interactivity of the three most used visualisation types

	Bar chart		Line chart		Map	
Level of interactivity	*N*	%	*N*	%	*N*	%
Level I	89	53.3	29	49.2	28	33.7
Level II	75	44.9	28	47.5	20	24.1
Level III	3	1.8	2	3.4	20	24.1
Level IV	0	0.0	0	0.0	10	12.0
Level V	0	0.0	0	0.0	5	6.0
Level VI	0	0.0	0	0.0	0	0.0
Total	167	100.0	59	100.0	83	100.0

visualisation type while being static in more than half of all cases (see Table 6). Contrarily, only one-third of all maps are static. Obviously, the characteristics of maps favour more interactive approaches than bar charts do. The third level of interactivity “allows the user to feel in control of the multimedia component representation, to view the component from different perspectives or in different sizes or actively navigate within it" (67). One can relate this to moving a map or zooming in and out of it. Almost one-quarter of all maps show these traits compared to 12 per cent being even more interactive. At the fourth level, the reader can set certain parameters in order to create new and non-predetermined data representations (68). Overall, maps occur as the most interactive visualisation type. After recoding the interactivity scale in order to check whether there is a relationship between the most used visualisation types (bar charts, line charts and maps) and the level of interactivity,[3] a significant relationship between these variables can be observed, $X^2(4, N = 309) = 86{,}60$, $p < 0.001$. This suggests that data journalists make use of genuine features of visualisation types, with maps being the most interactive format. Bar and line charts can also be interactive but apparently do not necessarily offer any substantial added value.

National, Governmental Data Sources

Those data sources that are visually represented were analysed. I do not consider sources that are mentioned in the text as these might include source types other than data sources—this would inevitably sway the categorisation of sources that are of interest within this study. Sources that are mentioned in the text are regarded provided that it is explicitly stated that they are visually represented and can be assigned to a specific chart. First, I examine data sources per article. In almost 30 per cent of all cases ($N = 73$), data journalists rely on more than one type of data source (see Table 7). They do so mostly by combining governmental data with data drawn from other sources. In 11.9 per cent of all cases, articles do not explicitly indicate whether data were used—this corresponds with the number of stories without visualisations; in 5 per cent of all cases ($N = 13$) I cannot determine the source of the data-set. The most used and clearly identifiable data providers are governmental bodies (17.6 per cent), followed by media and press releases (11.1 per cent) and research institutes (8.6 per cent). The combination of data-sets that leads to data sources coded as "Mixed" shows that data journalists apparently follow the two-source rule and try to balance different perspectives in their reporting. Overall, 24.6 per cent of all data-driven stories rely on at least one governmental source.

Examining data sources per article, the number of data sources each article employs was counted. There are some outliers and extreme values—for instance, one article

TABLE 7
Data sources per article

Source	*N*	% of total
Governmental	43	17.6
Mixed	73	29.9
Research	21	8.6
Media (release)	27	11.1
NGO	7	2.9
International organisation	12	4.9
Not known	13	5.3
Own data	11	4.5
National organisation	8	3.3
No data indicated	29	11.9
Total	244	100.0

indicates 25 sources—due to stories that make extensive use of various data (see Figure 2). After eliminating these extreme values, I can determine a mean value of 2.4 (SD = 2.4) sources per article. I can see that particularly stories published by *Zeit Online* show a wide variance in the number of sources used compared to other outlets.

Second, I regard data visualisations as coding units to get more granular information about the sources used per article: the 673 sources can be attributed to 510 visualisations. Almost three-quarters of all visualisations are based on one data source (*N* = 382); this leads to a mean of 1.3 sources per visualisation. In 7.1 per cent of all cases, authors do not explicitly indicate on which the sources their visualisations are based. On the contrary, 92.9 per cent of all visualisations can be clearly sourced since the names of respective data providers and/or links to the data-sets are appended to the visualisations. This leads to the assumption that—regarding sourcing—most data-driven stories are very thoroughly crafted.

According to Weinacht and Spiller (2014, 418–420), data journalists stated that providing used data-sets to readers as defined by open data is an essential characteristic of data journalism. In order to put this into practice, data journalists should provide a link to the source so that readers can access the data-sets or provide direct download links for the data-sets. A mere textual reference to a data-set would not qualify. Each article was scanned to determine whether this option was offered. The data suggest that, in general, the authors do not provide the data they have used. In more than half of all cases data-sets are not made available to readers. Some data are only partially available;

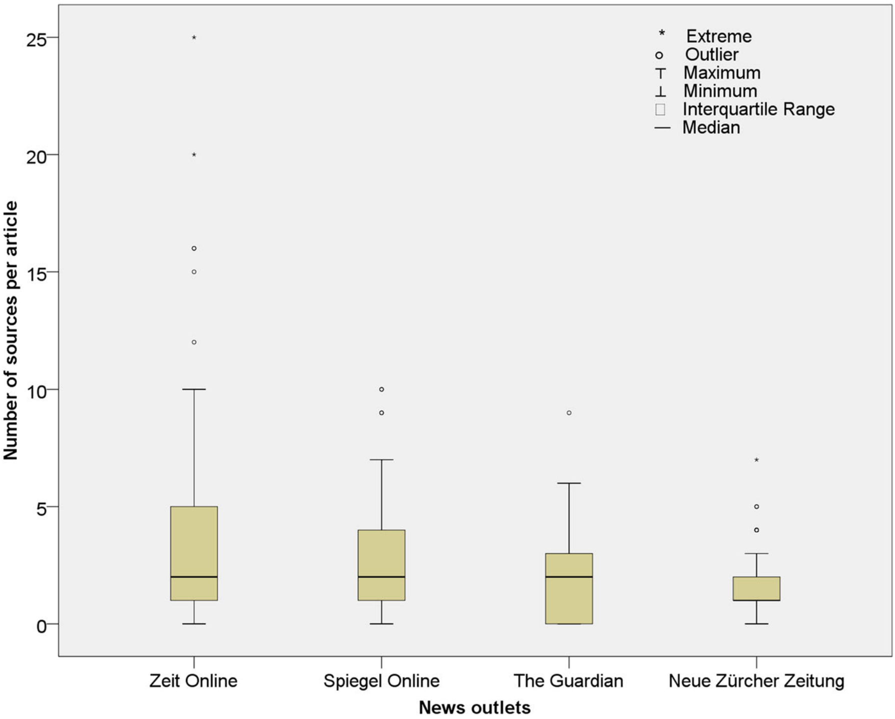

FIGURE 2
Number of sources per article grouped by news outlets

oftentimes if more than one data-set is used within an article then readers cannot access all of them. As a result, only 29.1 per cent of all data-driven stories (*N* = 244) provide all used data-sets to the reader.

Governmental sources comprise data providers such as governmental bodies, official statistical bureaus and, in general, public data released by governments. This also includes data provided by intergovernmental unions such as the European Union and its bodies.

The data suggest that almost a quarter of all data sources can be categorised as governmental, thereby constituting the most used source overall (see Table 8). In 17.1 per cent of all cases, visualisations are based on multiple sources ruling out an unambiguous categorisation. Yet again, in more than half of these cases (*N* = 45) governmental sources are implied. As a result, the overall employment of data drawn from governmental sources is even higher. The constraints of a quantitative content analysis, however, deny a detailed analysis of mixed data sources. Remarkably, 11 per cent of all sources are provided by other media outlets or are media releases. Therefore, I assume slight effects of co-orientation based on the reuse of already published data. Similarly, media releases pose other sources that contain pre-processed data. Contrary to expectations, data from international organisations such as the United Nations and its chapters or the World Bank are comparatively rare, amounting to 7.8 per cent. Genuine data collected by journalists themselves are only observable in 6.5 per cent (*N* = 33) of all cases.

To gain a more detailed insight into data sources, they were inspected regarding the regional origin and each source was coded according to the residence of its respective data provider. Germany, the United Kingdom and Switzerland are of particular interest as they are home to the four sampled news outlets (see Table 9). In more than two-thirds of all cases, *The Guardian* relies on domestic data sources (*N* = 82). Also, *Zeit Online* (*N* = 93) and *Spiegel Online* (*N* = 95) predominantly use national sources. Notably, *Neue Zürcher Zeitung* draws on Swiss data sources in 46.5 per cent of cases compared to 39.9 per cent of their sources being multinational, i.e. provided by multinational non-governmental organisations (NGOs) and intergovernmental organisations such as the United Nations or World Bank, or the European Union.

TABLE 8
Data sources per visualisation

			Zeit Online		***Spiegel Online***		***The Guardian***		***Neue Zürcher Zeitung***	
Source	***N***	**% of total**	***N***	**%**	***N***	**%**	***N***	**%**	***N***	**%**
Governmental	124	24.3	19	13.4	60	36.8	21	17.6	24	27.9
Mixed	87	17.1	47	33.1	22	13.5	2	1.7	16	18.6
Research	66	12.9	21	14.8	26	16.0	13	10.9	6	7.0
Media (release)	56	11.0	9	6.3	11	6.7	25	21.0	11	12.8
NGO	41	8.0	13	9.2	4	2.5	18	15.1	6	7.0
International organisation	40	7.8	3	2.1	20	12.3	8	6.7	9	10.5
Not known	36	7.1	16	11.3	11	6.7	1	0.8	8	9.3
Own data	33	6.5	12	8.5	5	3.1	14	11.8	2	2.3
National organisation	27	5.3	2	1.4	4	2.5	17	14.3	4	4.7
Total	510	100.0	142	100.0	163	100.0	119	100.0	86	100.0

TABLE 9
International and national data sources

	Zeit Online		*Spiegel Online*		*The Guardian*		*Neue Zürcher Zeitung*	
Source	*N*	%	*N*	%	*N*	%	*N*	%
Germany	93	65.5	95	58.3	0	0.0	2	2.3
United Kingdom	7	4.9	8	4.9	82	68.9	0	0.0
Switzerland	1	0.7	0	0.0	0	0.0	40	46.5
Other	14	9.9	13	8.0	20	16.8	3	3.5
Multinational	16	11.3	41	25.2	14	11.8	34	39.9
Not known	11	7.7	6	3.7	3	2.5	7	8.1
Total	142	100.0	163	100.0	119	100.0	86	100.0

These findings correspond with the results drawn from analysing references to foreign news in the articles. With only 45.9 per cent of all stories being purely domestic news, I rate *Neue Zürcher Zeitung* as the most international news site, followed by *Spiegel Online* (55.7 per cent), *The Guardian* (62.3 per cent) and *Zeit Online* (65.6 per cent). Regarding both the internationality of used data sources and subject matters presented on a textual level, *Neue Zürcher Zeitung* turns out to be spearheading the present sample regarding internationality.

Form and Content

Looking into the form and content of data-driven stories, variables that could make the authors' approaches measureable to some extent were identified. I can extract some meaningful information from the text level that is related to the journalists' underlying work—dependent on the style and composition of each individual article.

Dealing with subject matters such as education, 18 out of 19 related articles published on *The Guardian* website extensively cover yearly results of GCSEs and A-level examinations in the United Kingdom. Without regard for this overrepresented topic, many political subject matters are being put on the agenda: elections ($N = 19$), migration ($N = 10$) or politicians and parties ($N = 5$). The results suggest that soft news topics such as entertainment ($N = 1$) or fashion ($N = 1$) are rare cases. This indicates that data journalists focus on hard news and current topics—certainly, this depends on the agenda of the analysed media outlets. The use of story formats underscores the assumption that data-driven articles are hard news predominantly held in a formal, serious tone. Across all news outlets, 57.4 per cent do not refer to foreign news or other countries at all, implying that most topics are related to current domestic news. The vast majority of stories can be identified as news reports (76.2 per cent, $N =$ 186), followed by features (15.2 per cent, $N = 37$). Comments ($N = 5$), portraits ($N = 5$) and other formats ($N = 11$) seem to be exceptional cases. Notably, the sample contains rather exotic, gamified variants of data journalism: quiz games that compare users' scores to underlying databases ($N = 3$).

Moreover, I analysed if there is a perceived recognisable conflict by examining whether there is a perceivable contrasting juxtaposition of protagonists and antagonists, or whether the authors address problems that explicitly relate to objects of investigation. I can label 77 per cent (*N* = 187) of all articles neutral as they do not bear any of these compositions; 23 per cent (*N* = 57) of the sampled stories clearly point out problems or wrongdoings that can be linked to objects of investigation. The juxtaposition of two parties or a critical stance on stakeholders, holding them accountable, etc., are just some indicators for investigative approaches. These features alone certainly cannot unambiguously identify investigative stories—being an essential aspect of investigative reporting, they are, however, a first hint. An active journalistic approach, i.e. own analyses of data, the use of sources that make exposures more likely, accessing these sources against resistance, a socially highly relevant subject matter and a distinctive designation of causes of a problem—qualitative methodologies might be better suited to examine these other characteristics to determine investigative stories.

Comparative Discussion

To discuss the results of this study, I will contrast the findings with those of previous analyses of data journalistic articles. The findings show that daily data journalism usually involves one journalist (60 per cent); one-quarter is edited by a team of two journalists. Stories submitted to the Global Editors Network Data Journalism Award are authored by over five people on average (Reimer and Loosen 2017, 8), a quarter of Canadian projects by two and more than a quarter by six or more people (Young, Hermida, and Fulda 2017, 7). In theory, collaboration is promoted as a characteristic of data journalism (Fink and Anderson 2015, 472; Karlsen and Stavelin 2014, 45; De Maeyer et al. 2015, 442; Borges-Rey 2016, 6)—seeing collaborative efforts realised by crediting stories to teams of more than two editors, apparently does not account for daily data journalism but for bigger projects. The current sample suggests that data-driven articles can predominantly cover politics topics (almost 40 per cent), with another quarter society and business topics. Reimer and Loosen (2017, 9) came to similar conclusions.[4] Exemplarily, they state that "[m]any pieces in the politics section deal with elections" (10), paralleling my findings that elections are one of the most present subject matters in this study. Although Weinacht and Spiller (2014) consider data journalism a cross-sectional task, I assume a political focus.

Regarding data visualisations, in more than three-quarters of all cases, daily data stories contain both discrete text components as well as visualisations, while text-only and visualisation-only are rare. The results indicate that these stories usually contain two visualisations. This shows that data visualisations are a distinctive and very central feature of data-driven stories. Every third visualisation is a bar chart, followed by maps and line charts. These three types offer different explanatory functions: whereas bar charts display comparisons of different categories and values, maps add a geographical framework by locating stories in their regional context, and line charts predominantly include a temporal dimension by showing developments over time. These results correlate with Reimer and Loosen's (2017, 15) study which found that visualisations widely aim at comparing values or presenting changes over time. Whereas maps are found in almost half of all best-practice data-driven stories (Young, Hermida, and Fulda 2017, 9; Reimer and Loosen 2017, 16) and in more than half of stories from Quebec (Tabary, Provost, and Trottier 2016, 78), this visualisation type only accounts for 16.3 per cent of the 510 graphics analysed in the current study. This implies that geospatial data are harder to acquire or to visualise. Young, Hermida, and Fulda (2017, 13), however, argue that creating maps with free tools such as Google Maps is easy and frequently used even though such elements may lead to readability and usability problems for certain stories. Still, I assume that bar charts remain the most accessible and easily employable data representations—particularly considering the time pressure of daily journalism practice. At the same time, I found that maps are the most interactive means of data visualisation. In her research, Knight (2015, 62) determined that most map elements are static ($N = 20$) while more dynamic maps ($N = 8$) predominantly occur in newspapers such as *The Guardian* or *The Independent*. While these results are therefore only partially comparable to the present findings as Knight's sample does not only comprise quality news outlets, she determined that roughly more than one-third of all visualisations are maps and thereby significantly less than presented in other studies (Knight 2015, 62). The study by Tandoc and Oh (2017, 1009) also shows deviating results as they determined tables (28.8 per cent) as the most used data visualisation while maps only account for 1.2 per cent. I assume these results are caused by an overrepresentation of stories in the sample that were published via the *The Guardian*'s original data blog that was primarily used as a data repository.

Aside from maps, half of all bar and line charts do not have interactive features whereas the other half offer basic interactive features. This complies with Reimer and Loosen (2017, 16), who observed static charts in more than half of their sample. Appelgren's (2017, 15) study on paternalism in data journalism shows that "projects with paternalistic elements often involve a low level of physical interactivity" as they take away control from readers and force them to follow the linearity of a story in order to present a certain journalistic angle. Expanding on this notion, particularly bar charts could be considered to have high paternalistic tendencies as well as one-third of all maps. As I used Schulmeister's (2003) taxonomy of interactivity—higher levels of interactivity allow more manipulation by the user—maps appear to be the least paternalistic elements of data-driven storytelling. In addition, I assume that also sources can have paternalism effects.

Every fourth visualisation explicitly indicates governmental bodies as the data source and every fourth story uses at least one governmental data source. Tandoc and Oh's (2017, 1009) results show that 30 per cent of *The Guardian*'s big data stories rely on governmental data. Aside from *Neue Zürcher Zeitung*, all news websites primarily rely on domestic sources. Other most frequently used sources are research institutions such as universities, market research institutes, press releases or other media outlets. This supports the thesis put forward by Tabary, Provost, and Trottier (2016, 75) that data journalism is dependent on "pre-processed public data" to some extent. In most cases, data are drawn from domestic data sources despite a supposedly easy access to international data sources through global data portals. Therefore, I assume a domestication of data-driven news in accordance with news outlets' agendas and related target audiences. This finding raises the question of to what extent the availability and accessibility of certain data-sets influence the objectives of news organisations' agendas. Considering this dominance of domestic and governmental sources, I recommend follow-up research on how a pre-processed, officially published data-set is being employed, whether it has been "re-published" without scrutiny, to what extent numbers and methodologies have been rechecked, and whether it is employed to support a thesis or whether it is an object of investigation itself. Furthermore, 11 per cent of all visualised data-sets are provided—and again pre-processed—by media outlets or through press releases, thus, I can assume at least slight effects of co-orientation based on the reuse of already journalistically published data. Original data collected by media outlets or journalists themselves is only observable in 6.5 per cent of all cases. Tandoc and Oh (2017, 1009) assume an increasing independence of governmental data through the availability of data published by international organisations or through own data collections—the present results, however, do not confirm these observations. The present findings strongly advocate Tabary, Provost, and Trottier's claim:

> In order to develop more meaningful and deep-digging data journalism for the future, journalists must control data collection (and, more generally, numerical proof protocols). If data journalism is understood to include original statistical analysis, building a database is a fundamental part of the process. Adopting the naïve positivist position (Desrosières, 2008) that "numbers speak for themselves" (p. 10) means giving to analysis performed by others and, in this case, relaying information created by the institutions in power. (Tabary, Provost, and Trottier 2016, 81)

Using original data is more common to best-practice data journalism, accounting for 20 per cent of all submitted stories (Reimer and Loosen 2017, 12). This discrepancy indicates that collecting and analysing one's own data is not (yet) feasible for daily data journalism as time and resources are limiting factors. A general lack of experience in social science methods or a sporadically missing socio-scientific education of journalists might be additional thinkable factors. Indeed, results of the Global Data Journalism Survey (Heravi 2017, 6) indicate shortcomings in social science education and training in data analysis or statistics and particularly in data science or coding. This concurs with Borges-Rey (2017, 9), who found that only a few data journalists and editors are familiar with Web scraping and thereby generating their own data.

Transparency is a central characteristic according to Coddington (2015, 337); also Weinacht and Spiller (2014, 418–420) consider providing readers with the data used to be a core characteristic of data journalism. The current findings show that in more than half of all stories data were not provided at all—in accord with Young, Hermida, and Fulda's (2017, 12) results.

Both analyses of acclaimed data-driven stories found that investigative formats (Young, Hermida, and Fulda 2017, 8) or watchdog journalism (Reimer and Loosen 2017, 10) oftentimes go hand in hand with data journalism. By examining the structure and story line of daily data stories, I assume that roughly every fourth article critically points out wrongdoings by holding respective antagonists such as companies or politicians accountable. In order to determine fully the investigative nature of a journalistic product, one would need to get further insights into processes that are not evident on a text level: the role of journalists, whether they encountered resistance or whether there had been attempts to deter authors from investigating, assessing the relevance of the subject matter and if there had been any consequences or repercussions due to the reporting; getting a comprehensive picture to determine an investigative story requires additional findings which cannot be drawn from a quantitative study.

Conclusions and Perspectives

In his case study, Borges-Rey concludes that data journalism in the United Kingdom

> has largely diversified into three forms of data journalism: (1) a daily, quick turnaround, generally visualised, brief form of data journalism; (2) an extensive, thoroughly researched, investigative form of data journalism; and (3) a light, editorialised, entertaining, often-humorous, gamified form of data journalism. (Borges-Rey 2016, 9)

These three streams provide suitable concepts for classifying data-driven articles. What this study adds are empirical results that allow elaboration of the frequency of appearance of these forms of data journalism. Whereas previous content analyses—primarily due to the composition of samples—convey the impression that data-driven stories are in-depth, long-form, oftentimes investigative and visually sophisticated projects, this research puts available results into perspective. Reimer and Loosen (2017, 22), discussing their methodology, state that "the analysed pieces are based on self-selection, and, second, they are likely not to represent 'everyday' data journalism". Clearly, this study might also not be representative of the population of data journalistic articles since big, investigative projects are simply rare and not necessarily being published via websites/landing pages from which the present sample was drawn. What this study achieves, however, is a detailed delineation of everyday data journalism.

Overall, daily data-driven stories can distinguish themselves from traditional journalistic products through a focus on data sources as primary sources and visualisations. Employed visualisations—despite certainly being a key characteristic of data journalism—oftentimes lack interactive features disregarding added values offered by digital and internet technologies, while employing obsolescent visualisation types. The findings suggest that data journalism is not a new holistic genre of journalism but a journalistic practice that, when employed, can enhance stories with visualisations while also enabling journalists to incorporate data sources as primary sources that previously may have widely been regarded as inoperative. At what point the mere availability of data influences or even ignites a journalistic investigation should be further discussed through qualitative design or through the monitoring of data-set releases and whether or not they are picked up by data journalists. This links to the two approaches of data-driven investigations as put forward by Parasie (2015, 373): a "hypothesis-driven approach" and a "data-driven approach".

After giving these insights into daily data journalism in Europe, I suggest comparative follow-up studies on daily data-driven stories in other regions, given that data journalism has already found its way into respective newsrooms at least to some extent. I hope that this research also motivates more detailed qualitative close-ups of selected data-driven projects combining qualitative and quantitative methodologies.

ACKNOWLEDGEMENTS

I would like to thank the anonymous reviewers for their constructive and detailed comments.

FUNDING

No grants or funding bodies supported this research.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

NOTES

1. The Global Editors Network Data Journalism Award, the Marketwired Data Journalism Award and the investigative data journalism award of the Online News Association.
2. Some of these landing pages might have changed. Reorganisations within the news outlets over the last year have led to new teams and new output channels.
3. Level I = not interactive; level II = interactive; levels III, IV, V and VI = highly interactive.
4. They found 48.6 per cent political topics, 34.6 per cent societal topics and 23.5 per cent business topics.

REFERENCES

Appelgren, Ester. 2017. “An Illusion of Interactivity. The Paternalistic Side of Data Journalism.” *Journalism Practice*. doi:10.1080/17512786.2017.1299032.

Appelgren, Ester, and Gunnar Nygren. 2014. “Data Journalism in Sweden: Introducing New Methods and Genres of Journalism Into ‘Old’ Organizations.” *Digital Journalism* 2 (3): 394–405. doi:10.1080/21670811.2014.884344.

Borges-Rey, Eddy. 2016. “Unravelling Data Journalism: A Study of Data Journalism Practice in British Newsrooms.” *Journalism Practice* 1–11. doi:10.1080/17512786.2016.115992.

Borges-Rey, Eddy. 2017. “Towards an Epistemology of Data Journalism in the Devolved Nations of the United Kingdom: Changes and Continuities in Materiality, Performativity and Reflexivity.” *Journalism*. doi:10.1177/1464884917693864.

Coddington, Mark. 2015. “Clarifying Journalism’s Quantitative Turn: A Typology for Evaluating Data Journalism, Computational Journalism, and Computer-assisted Reporting.” *Digital Journalism* 3 (3): 331–348.

De Maeyer, Juliette, Manon Libert, David Domingo, François Heinderyckx, and Florence Le Cam. 2015. “Waiting for Data Journalism: A Qualitative Assessment of the Anecdotal Take-up of Data Journalism in French-speaking Belgium.” *Digital Journalism* 3 (3): 432–446. doi:10.1080/21670811.2014.976415.

Desrosières, Alain. 2008. *Pour une sociologie historique de la qunatification. L'Argument statistique I*. Paris: Presses de l'Ecole des mines.

Fink, Katherine, and C. W. Anderson. 2015. “Data Journalism in the United States. Beyond the ‘Usual Suspects’.” *Journalism Studies* 16 (4): 467–481.

Gynnild, Astrid. 2014. “Journalism Innovation Leads to Innovation Journalism: The Impact of Computational Exploration on Changing Mindsets.” *Journalism: Theory, Practice & Criticism* 15 (6): 713–730. doi:10.1177/1464884913486393.

Heravi, Bahareh. 2017. “The State of Data Journalism Globally.” In *Proceedings of the European Data and Computational Journalism Conference*, edited by Tomas Petricek, Bahareh R. Heravi, Jennifer A. Stark, Nicholas Diakopoulos, Martin J. Chorley, Glyn Mottershead, Marcel Broersma, and Marc Esteve de Valle. Dublin: University College Dublin. http://researchrepository.ucd.ie/handle/10197/8634.

Karlsen, Joakim, and Eirik Stavelin. 2014. “Computational Journalism in Norwegian Newsrooms.” *Journalism Practice* 8 (1): 34–48. doi:10.1080/17512786.2013.813190.

Kirk, Andy. 2016. *Data Visualisation*: *A Handbook for Data Driven Design*. London: Sage.

Knight, Megan. 2015. “Data Journalism in the UK: A Preliminary Analysis of Form and Content.” *Journal of Media Practice* 16 (1): 55–72. doi:10.1080/14682753.2015.1015801.

Parasie, Sylvain. 2015. “Data-driven Revelation? Epistemological Tensions in Investigative Journalism in the Age of ‘Big Data’.” *Digital Journalism* 3 (3): 363–380. doi:10.1080/21670811.2014.976408.

Parasie, Sylvain, and Eric Dagiral. 2013. “Data-driven Journalism and the Public Good: ‘Computer-assisted-reporters’ and “Programmer-journalists” in Chicago.” *New Media & Society* 15 (6): 1–19. doi:10.1177/1461444812463345.

Reimer, Julius, and Wiebke Loosen. 2017. “Data Journalism at its Finest: A Longitudinal Analysis of the Characteristics of Award-nominated Data Journalism Projects.” In *News, Numbers, and Public Opinion in a Data-driven World*, edited by An Nguyen (Forthcoming). New York: Bloomsbury.

Roam, Dan. 2009. *The Back of the Napkin*: *Solving Problems and Selling Ideas with Pictures*. Expanded ed. New York: Penguin Books.

Schulmeister, Rolf. 2003. “Taxonomy of Multimedia Component Interactivity. A Contribution to the Current Metadata Debate.” *New Media in Education - Special Issue of Studies in Communication Sciences* 3 (1): 61–80.

Tabary, Constance, Anne-Marie Provost, and Alexandre Trottier. 2016. “Data Journalism’s Actors, Practices and Skills: A Case Study from Quebec.” *Journalism: Theory, Practice & Criticism* 17 (1): 66–84. doi:10.1177/1464884915593245.

Tandoc, Edson C., and Soo-Kwang Oh. 2017. “Small Departures, Big Continuities? Norms, Values, and Routines in The Guardian’s Big Data Journalism.” *Journalism Studies* 18 (8): 997–1015. doi:10.1080/1461670X.2015.1104260.

Uskali, Turo I., and Heikki Kuutti. 2015. “Models and Streams of Data Journalism.” *The Journal of Media Innovations* 2 (1): 77–88.

Weinacht, Stefan, and Ralf Spiller. 2014. “Datenjournalismus in Deutschland. Eine explorative Untersuchung zu Rollenbildern von Datenjournalisten.” *Publizistik* 59 (4): 411–433.

Young, Mary Lynn, Alfred Hermida, and Johanna Fulda. 2017. “What Makes for Great Data Journalism? A Content Analysis of Data Journalism Awards Finalists 2012–2015.” *Journalism Practice* 141: 1–21. doi:10.1080/17512786.2016.1270171.

Florian Stalph, Centre for Media and Communication, University of Passau, Germany. E-mail: florian.stalph@uni-passau.de. Web: http://www.phil.uni-passau.de/journalistik/team/florian-stalph/