Wiebke Loosen / Julius Reimer / Fenja De Silva-Schmidt

Data-Driven Reporting – an On-Going (R)Evolution?

A Longitudinal Analysis of Projects Nominated for the Data Journalism Awards 2013–2015

Arbeitspapiere des Hans-Bredow-Instituts Nr. 41 | Mai 2017

HANS-BREDOW-INSTITUT
für Medienforschung an der Universität Hamburg

Wiebke Loosen / Julius Reimer / Fenja De Silva-Schmidt (2017): Data-Driven Reporting – an On-Going (R)Evolution? A Longitudinal Analysis of Projects Nominated for the Data Journalism Awards 2013–2015. Hamburg: Verlag Hans-Bredow-Institut, Mai 2017 (Arbeitspapiere des Hans-Bredow-Instituts Nr. 41)

ISSN 1435-9413
ISBN 978-3-87296-141-9

Hans-Bredow-Institut for Media Research at Universität Hamburg

The research conducted by the Hans-Bredow-Institut focuses on mediated public communication: Today, the different types of mass media shape people's everyday life, politics, the economy as well as culture to a greater extent than ever before. Understanding the underlying determinants, assessing future opportunities and risks, and providing orientation for the actors involved – this is the main focus of the Institute's research. The Institute, founded as an independent non-profit organisation in 1950, emphasises its role as an independent observer and combines sociological, legal, economic and pedagogical approaches, because it strongly believes that contemporary problems of media development call for an interdisciplinary and a cross-national comparative perspective. Cooperation and constant exchange with the different actors in the media sphere are basic values for the Institute's research. Particularly important is the transfer of results via publications and conferences in direction of business practice, politics and the public sphere.

The authors

PD Dr. Wiebke Loosen (w.loosen@hans-bredow-institut.de) and Julius Reimer (j.reimer@hans-bredow-institut.de) work at the Hans-Bredow-Institut for Media Research at the University of Hamburg, Hamburg, Germany. Fenja De Silva-Schmidt (fenja.schmidt@wiso.uni-hamburg.de) works at the Faculty of Economic and Social Sciences, Universität Hamburg, Hamburg, Germany

Publisher

Hans-Bredow-Institut für Medienforschung an der Universität Hamburg
Rothenbaumchaussee 36
20148 Hamburg / Germany
Tel.: (+49 40) 450 217-12
E-Mail: info@hans-bredow-institut.de

Contents

Abstract

Wiebke Loosen / Julius Reimer / Fenja De Silva-Schmidt (2017): Data-Driven Reporting – an On-Going (R)Evolution? A Longitudinal Analysis of Projects Nominated for the Data Journalism Awards 2013–2015

The emergence of data-driven journalism (DDJ) can be understood as journalism's response to the datafication of society. We retrace the development of this emerging reporting style by looking at what may be considered the gold-standard in data-driven reporting: projects that were nominated for the Data Journalism Awards (DJA), a prize issued annually by the Global Editors Network. Using a content analysis of the nominees from 2013 to 2015 (n = 179) we examine if and how, among other aspects, data sources and types, visualisation strategies, interactive features, topics, and types of nominated media outlets have changed over the years. Results suggest, for instance, that the set of structural elements data-driven pieces are built upon remains rather stable, that data journalism is increasingly personnel intensive and progressively spreading around the globe, and that journalists, while still concentrating on data from official institutions, are increasingly looking to unofficial data sources for their stories.

Keywords: data, data journalism, data-driven journalism, content analysis, Data Journalism Awards, reporting style, presentation form, visualisation

Introduction

The emergence of data-driven journalism (DDJ) can be understood as journalism's response to the datafication of society. In fact, the phenomena of 'big data' and an increasingly data-driven society are doubly relevant for journalism: Firstly, it is a topic worth covering so that the related developments and their consequences can be understood in context and public debate about them can be encouraged. Secondly, the 'quantitative turn' (Coddington, 2015) has already begun to affect news production itself and has given rise to novel ways of identifying and telling stories (Lewis and Usher, 2014): As a consequence, we are witnessing the emergence of a new journalistic sub-field often described as 'computational journalism' (Karlsen and Stavelin, 2014) or 'data-driven journalism' (Borges-Rey, 2016: 840), acronymously known as, 'DDJ'.

The extensive attention that practitioners pay to DDJ has also fuelled 'an explosion in data journalism-oriented scholarship' (Fink and Anderson, 2015: 476). This research is based on case studies, cursory observations, and/or samples that are limited in spatio-temporal terms. We aim to complement this body of work with a longitudinal, international study of what may be considered the gold-standard among practitioners: projects nominated for the Data Journalism Awards (DJA) from 2013 to 2015. Through a content analysis of these pieces we look at how this new reporting style and its key components (e.g. data sources and types, visualisation strategies, interactive features) develop over time. Three years may appear to be a rather short period for a longitudinal analysis, but given the 'rapidly changing nature' (Royal and Blasingame, 2015: 41) of data journalism, we expect it to be long enough to shed an initial light on developments in the field.

Previous work and research objectives

Scholarship on DDJ has been dominated by three particular areas of study: Firstly, researchers have tended to focus on the actors involved in the production of data journalism: Data journalists in Belgium (De Maeyer et al., 2015), Germany (Weinacht and Spiller, 2014), Norway (Karlsen and Stavelin, 2014), Sweden (Appelgren and Nygren, 2014), the United Kingdom (Borges-Rey, 2016) and the United States (Boyles and Meyer, 2016; Fink and Anderson, 2015; Parasie, 2015; Parasie and Dagiral, 2013) have been interviewed and observed with regard to their (journalistic) self-understanding and (the organisation of) their work in the newsroom.

Secondly, scholars have tried to clarify what data journalism actually is and how it is similar to and different from investigative journalism, computer-assisted reporting, computational journalism, etc. (e.g. Coddington, 2015; Fink and Anderson, 2015; Royal and Blasingame, 2015). Against this backdrop and based on cursory observations of the field and example projects, scholars have more-or-less agreed on the following key characteristics of DDJ:

- It usually builds on (large) sets of (digital) quantitative data as 'raw material' that is subjected to some form of (statistical) analysis in order to identify and tell stories (Coddington, 2015; Royal and Blasingame, 2015);
- its results 'often need visualization' (Gray et al., 2012: n.p.), i.e. they are presented in the form of maps, bar charts and other graphics (Royal and Blasingame, 2015);
- it is 'characterised by its participatory openness' (Coddington, 2015: 337) and 'so-called crowdsourcing' (Appelgren and Nygren, 2014: 394) in that users help with collecting, analysing or interpreting the data (cf. also Borges-Rey, 2016; Boyles and Meyer, 2016; Karlsen and Stavelin, 2014);
- it is regularly related to an open data and open source approach meaning that it is regarded as a quality criterion of DDJ that journalists also publish the raw data a story is built upon (Gray et al., 2012).

A third strand of research has analysed the actual data-driven content that is produced. These studies focus on the above-mentioned elements and affirm their status as key characteristics of a data-journalistic reporting style. However, in spatial or temporal terms, their samples are rather limited: Parasie and Dagiral's (2013) study refers to pieces from one Chicago outlet published before March 2011; Knight (2015) analyses articles published in fifteen UK newspapers over a two-week period in

2013; Tandoc and Oh (2015) turn to 260 stories published in *The Guardian*'s *Datablog* between 2009 and 2015; Tabary et al. (2016) examine projects produced between 2011 and 2013 by six Québécois media outlets. In the results section, we will draw on the results of these content analyses, wherever applicable, to put our own findings into perspective.

Against this backdrop, we can conclude that present research reflects:

- a certain knowledge about DDJ's actors, their self-image (as journalists), and selective knowledge of their integration in- and outside of established newsrooms,
- various attempts to define DDJ as a distinguishable reporting style that revolve around some apparent core characteristics, and
- initial studies that empirically analyse DDJ products, but are restricted in their scope.

This leads to the paradoxical situation that, while practitioners as well as academics frequently associate data-driven reporting with the *future* of journalism (Gray et al., 2012; Lewis and Usher, 2014), we know little about how it is *currently* developing, if at all.

We aim to take a first step towards closing this gap by conducting a longitudinal analysis on a broader geographical scale that advances our understanding of how DDJ as 'an *emerging* form of storytelling' (Appelgren and Nygren, 2014: 394; *emphasis added*) is currently evolving over time. For research into a relatively new and constantly changing phenomenon like this, a mainly descriptive approach offers an appropriate starting point as it can lay down systematic empirical groundwork needed for further analyses.

In pursuit of this aim we have clarified the following research questions:

RQ 1: What structural elements and forms of presentation are data-driven pieces composed of and how is this composition evolving?

RQ 2: How are the topics covered in data-driven projects changing over time?

RQ 3: How is the field of actors producing data journalism (media organisations, in-house teams, external partners) developing?

As presenting all our empirical findings with the same level of detail is beyond the scope of this paper, we will mainly focus on RQ1 as it addresses the assumed core characteristics of data-driven pieces.

Method

We attempt to answer our research questions by conducting a standardised content analysis (e.g. Krippendorff, 2013) of data-journalistic pieces from a three year period and on a broad geographical scale.

Since data journalism is such a 'diffuse term' (Fink and Anderson, 2015: 470), it is difficult, or rather preconditional, to identify respective pieces for a content analysis. That is why we have decided on an inductive and pragmatic approach that avoids starting with either too narrow or too broad a definition of what counts as data journalism. As such, our sample (see table 1) consists of pieces nominated for the *Data Journalism Awards* (DJA) – a prize awarded annually by the *Global Editors Network*[1] – in 2013, 2014 and 2015. In other words, *the field itself* considers these projects as data journalism and believes that they represent significant examples of this reporting style. Similar approaches to sampling have already proven useful for analysing particular genres and aspects of storytelling (Lanosga, 2014; Wahl-Jorgensen, 2013). Additionally, we chose to analyse DJA-nominees because of their likely impact on the field in general as "[t]hrough transferring the capital that awards bring, certain sectors of the media can […] shape professional standards" (Jenkins and Volz 2016: 16; authors' italics).

However, we must take into account that our sample is doubly biased: First, the analysed pieces are based on self-selection as any data journalist can submit her/his work to be considered for nomination by the organising committee. Second, nominees for a data journalism award are not likely to represent 'everyday' data journalism. The field has already diversified, and our sample very likely consists of 'an extensive, thoroughly researched, investigative form of data journalism' which, as Borges-Rey (2016: 841) found out for the UK, is quite distinguishable from 'a daily, quick turnaround, generally visualised, brief form of data journalism' (cf. also De Maeyer et al., 2015; Fink and Anderson, 2015). This is especially true because research has shown that awards tend to favour nominees that already enjoy a high status in the field (Jenkins and Volz, 2016).

1 Cf. http://www.globaleditorsnetwork.org/about-us/ (accessed 20 December 2016).

Table 1: Dataset overview[2]

		2013	2014	2015	Total
Submissions	Freq	>300[I]	520	482	>1,302
Nominated projects	Freq	72	75	78	225
	% of submissions	<24.0	14.4	16.2	<17.3
Projects suited for analysis	Freq	56	64	59	**179**
	% of nominees	77.8	85.3	75.6	79.6
	% of projects analysed	31.3	35.8	33.0	100.0
Award-winning projects	Freq	6	9	13	28
	% of projects analysed	10.7	14.1	22.0	15.6

[I] The GEN does not specify the number of submissions for 2013, but only states that 'more than 300 entries' had been submitted (http://www.globaleditorsnetwork.org/programmes/dja; accessed February 17, 2014).

While three years may not be a long time for a longitudinal perspective, in a dynamic field such as data journalism, where changes may occur from one year to the next, it might be worthwhile to track developments over short(er) time spans. Nonetheless, we interpret year-to-year differences with caution because of the long production times that often characterise data-driven projects: The production of some 2015 nominees, for example, may overlap with that of some 2014 pieces and because of this, the lines between the years sometimes become blurred.

One advantage of our sample is that it allows us to identify differences between those data journalism pieces that were only nominated and those that actually won an award, i.e. between what is considered high quality and the best practice in the field.

Our codebook comprises, amongst others, the (presumed) key characteristics of DDJ listed in the review of previous work. Most variables and their values, that is the different forms a variable can take, were developed inductively in 2013, based on an explorative analysis of a subsample from that

2 If a nomination referred to a media outlet as a whole and not to a specific project, the case was excluded from the analysis as our unit of analysis is a single data-driven piece. A list of (and links to) all projects nominated for a DJA in 2013, 2014 and 2015 is available on: http://community.globaleditorsnetwork.org/projects_ by_global_event/744 (accessed October 14, 2015).

year. Some categories were inspired by Parasie and Dagiral's (2013) study, the only content analysis of data-driven pieces available at the time; others were suggested by fellow researcher Julian Ausserhofer and data journalist Lorenz Matzat. A pretest was conducted with two coders and a subsample of ten percent of cases. All variables reached an intercoder reliability coefficient (Holsti or Krippendorff's Alpha) equivalent to or higher than 0.7 which is generally considered sufficient for exploratory research (Krippendorff, 2013).

The final codebook contains twenty-eight categories which can be grouped roughly into five dimensions (see table 2); with the presentation of the results we will provide deeper insights into most of the variable categories. It is notable that over the years we did not have to change or add variables or values to capture new kinds of data, visualisations or other elements, the only exception being audio files which were used in projects from 2015.

Table 2: Dimensions and variables of the codebook[3]

Dimension	Variables
authorship	medium; type of medium; external partners; number of people involved mentioned by name
story properties	headline; topic; reference to a specific event[I]; question(s) posed to data; number of related articles[II]; length of article; language; winner of DJA
data	data source(s); type(s) of data source(s); access to data; kind of data; additional information on data[I]; geographical reference; changeability of dataset[III]; time period covered; unit of analysis
analysis and journalistic editing of content	personalised case example[IV]; call for public intervention or criticism[II]; purpose of data analysis[V]; visualisation
interactive features	interactive functions; online access to the database[II]; opportunities for communication

I Suggested by data journalist Lorenz Matzat.
II Adopted from Parasie and Dagiral (2013: 5–14).
III Suggested by (data) journalism researcher Julian Ausserhofer.
IV Inspired by Holtermann (2011)
V Inspired by Gray et al. (2012: n.p.)

3 We will provide the complete codebook on request.

Results

The presentation of the results addresses the research questions in reverse order: It starts by shortly looking at the actors producing DDJ (*RQ3*) and the topics they cover (*RQ2*) as well as some formal story elements. These figures provide the background against which we will then present the results in terms of DDJ's 'key characteristics' as suggested by the literature: data-drivenness, visualisation, and interactivity (*RQ1*). Wherever indicated, we will also refer to the results of the above-mentioned, previous content analyses, actor and case studies to put our findings into perspective.

Actors producing DDJ

Over the three years we looked at, newspapers represent by far the largest – and most consistently expanding – group among all nominees as well as among the award winners (2013: 41.1%; 2015: 47.5%; total: 44.1%). Another important group are investigative journalism organisations such as *Pro Publica* and *The International Consortium of Investigative Journalists* (ICIJ) (total: 19.0%). Print magazines (8.9%), public and private broadcasters (5.6 and 2.8%), native online media, news agencies and non-journalistic organisations (5.0%, respectively), university media (3.4%) and other types of authors (1.1%) are represented to much lesser extents. This leaning towards print media and investigative organisations likely reflects the inherent bias of awards towards established, high-profile actors (Jenkins and Volz, 2016). This assumption is supported by the fact that those players often referred to as prime examples in the literature (e.g., Anderson, 2013; Coddington, 2015) have also been nominated most often for a DJA: *ProPublica* (16 cases), *The New York Times* (10 cases), the *ICIJ* and US magazine *Mother Jones* (both nine cases), *The Guardian* (12 cases). However, we also find less well known examples in the sample like the Argentinian newspaper *La Nación* (eight cases). An explanation for the predominance of newspapers might be that their institutional imperatives place a greater value on submitting work for major awards than more upstart outlets.

Our results illustrate that data journalism is usually a collaborative effort. In cases where the project contained a byline (n = 154), on average just over five individuals are named as authors or contributors (M = 5.16). This is probably due to the division of labour into data analysis, visualisation, and writing which several studies found to be common in the field (Tabary et al., 2016; Weinacht and Spiller, 2014). In this context, De Maeyer et al. (2015: 440–441) differentiate between '"ordinary" data journalism' that "is manageable by one individual [and] can be done on a daily basis" and '"thorough" data journalism' that is produced by teams with "a range of skills". Similar distinctions are made by Borges-

Rey (2016) as well as Fink and Anderson (2015). Furthermore, data journalism seems to have become progressively more personnel-intensive, since the average number of people involved in its production increased from 4.13 in 2013 to 5.67 in 2015.

According to the project descriptions that the authors provide when submitting their work to the DJA, nearly a third (31.3%) of all projects have been realised in association with external partners either contributing to the analysis or designing visualisations. In 2015, however, the share of projects that were realised with external partners is substantially lower than in the previous two years (2013: 33.9%, 2014: 35.9%, 2015: 23.7%). This might be a consequence of an increase in the recruitment of in-house qualified personnel and/or the result of intensified training within news organisations. Boyles and Meyer (2016: 949), for instance, found, through in-depth interviews with data journalists at American newspapers, that in 'many cases, data journalists were also relied upon to retrain other newsroom staff in new approaches to newsgathering—particularly database tools and code literacy'.

Nearly half of the nominees come from the United States (48.6%), followed at a distance by Great Britain (12.8%) and Germany (7.3%). Again, we must assume some sort of bias here: Firstly, data journalism has a longer history in English speaking countries and secondly, American news organisations may be more likely to submit their works to the DJA. However, the number of countries represented by the nominees grew with each year, amounting to twenty-seven countries from all five continents. This suggests that data journalism is increasingly spreading around the globe.

Not only does DDJ seem to be experiencing a global spread of its own accord, it appears that journalists are increasingly keen on appealing to an international audience as the share of bi- or multilingual (15.6%) projects is second only to purely English-language projects (67.0%) and increased from 14.3 to 18.6 percent. In most of these cases, the project is published in English and in the medium's native language.

Topics and formal story elements

Almost half of the analysed pieces cover a political topic (48.6%; multiple coding possible), a third deal with societal issues (census results, crime reports, etc.; 34.6%), nearly a quarter focus on business and the economy (23.5%) and more than a fifth are concerned with health and science (21.2%). Education, sports and culture attract little coverage (2.2% to 6.1%). Furthermore, data-driven stories appear to

have a clear thematic focus since nearly two-thirds of cases deal with only one category of topic (65.9%).[4]

Data-driven stories in the politics section deal, for instance, with elections, which tend to generate vast amounts of quantitative data. In some cases, political data-driven stories distinctly take on a watchdog role and check the validity of politicians' statements based on statistical data. In general, data journalism often assumes a critical position, since we found elements of criticism (e.g. on the police's wrongful confiscation methods) or even calls for public intervention (e.g. with respect to carbon emissions) in half of the pieces analysed (49.2%). This share is more or less stable over the three years and considerably higher among the award-winners (60.7% vs. 47.0%).

Data sets, sources, and analysis

The data journalism we analysed relied, to a large extent, on geodata (44.4%), financial data (43.3%), and measured values gathered by sensors or with measuring tools (42.1%; e.g. aircraft noise, train speeds and carbon emissions) (see table 3). While this last category has gained prominence over the years, award-winning projects are based on this kind of data to a below average extent (28.6%). Another type of data frequently analysed is sociodemographic data (32.6%). Two types of data used more than average in award-winning pieces are financial and personal data – i.e. information which can be attributed to individual persons; this is the only significant difference between award-winners and non-winners here. Metadata (i.e. 'data about data', for example, information about individual instances of application use) and data from polls and surveys are used least frequently.

4 Due to the very different scopes and samples of the above-mentioned content and analytical studies on data journalism, it is difficult to compare results. If we do so, nonetheless, our findings are, by and large, confirmed: from the considerable shares of political (Tandoc and Oh, 2015), societal (Knight, 2015) and business issues (Parasie and Dagiral, 2013) to the small proportions of stories on health and science, sports and culture (Knight, 2015; Tandoc and Oh, 2015).

Table 3: Kind of data (multiple coding possible)

%	**2013** (n = 55)	**2014** (n = 64)	**2015** (n = 59)	**Not awarded (2013-2015)** (n = 151)	**Awarded (2013-2015)** (n = 28)	**Total** (n = 179)
Geo data	47.3	39.1	47.5	46.0	35.7	44.4
Financial data	45.5	45.3	39.0	41.3	53.6	43.3
Measured values	34.5	43.8	47.5	44.7	28.6	42.1
Sociodemogr. data	38.2	25.0	35.6	32.0	35.7	32.6
Personal data	21.8	32.8	32.2	26.0[I]	46.4[I]	29.2
Metadata	12.7	20.3	13.6	16.0	14.3	15.7
Poll ratings / survey data	14.5	10.9	20.3	17.3	3.6	15.2
Other data	-	-	-	0.7	3.6	1.1

[I] Fisher's exact test: $p < .05$.

As expected, however, some kinds of data are used significantly more often in pieces dealing with particular topics. For instance, the above-mentioned information from polls/surveys is included significantly more often in political stories than in non-political ones (23.0%, $n = 87$, vs. 7.6%; $n = 92$; Fisher's exact test: $p < .01$). Economic and business pieces draw on financial data significantly more often than other stories (83.3%, $n = 42$, vs. 30.7%, $n = 137$; Fisher's exact test: $p < .001$). In turn, work on societal topics is significantly more likely than non-societal coverage to contain sociodemographic information (56.5%, $n = 62$, vs. 19.7%, $n = 117$; Fisher's exact test: $p < .001$) while measured values appear significantly more often in pieces that deal with health or science (78.9%, $n = 38$, vs. 31.9%, $n = 141$; Fisher's exact test: $p < .001$).

Only about a quarter of the pieces rely on only one type of data (24.6%) while most stories refer to two (40.8%) or three (24.0%) different kinds of information. Furthermore, the average number of different kinds of data used has grown slightly over the years (2013: $M = 2.14$, $SD = 0.96$; 2014: $M = 2.17$, $SD = 0.99$; 2015: $M = 2.36$, $SD = 1.05$). Most frequently, geodata was combined with measured values (21.2%; e.g. radiation levels or noise exposure), sociodemographic information (17.9%) or financial data (16.8%).

It is considered a quality criterion in data journalism that data sources should be cited (Gray et al., 2012); yet, 6.1 percent of the articles we surveyed did not indicate where they got their data from (see table 4). However, this is not the case for any of the award-winning pieces.[5]

By far, most pieces in our sample use data from official institutions like Eurostat and other statistical offices and ministries (68.2%). This reflects Tabary et al.'s (2016: 75) finding that data journalism exhibits a 'dependency on pre-processed public data'. The second largest group consists of pieces that use data from other non-commercial organisations including universities, research institutes and NGOs (44.1%). Roughly twenty percent analyse data that the respective media organisation collected itself, e.g. through a survey or by searching its own archives ('own source'). This share is much larger than the seven percent in Knight's (2015) analysis, but comparable to that found by Tandoc and Oh (2015) in their study of *The Guardian*'s *Datablog*. A comparison between years shows that basing stories on one's own data, after a drop in 2014, is on the rise again and the share of pieces that report data from private companies has grown consistently. This increase over the years, and the fact that researching exclusive data is thriving again in 2015, suggests that data journalists are looking to additional data sources for their stories instead of only relying on official sources. This assumption is supported by the fact that the average number of *different* types of sources referred to in a data-driven piece has risen from 1.40 in 2013 (*SD* = 0.66) to 1.68 in 2014 (*SD* = 0.63) and 1.67 in 2015 (*SD* = 0.80).[6]

Table 4: Type of data source (multiple coding possible)

%	**2013** (n=56)	**2014** (n=64)	**2015** (n=59)	**Not awarded (2013-2015)** (n=151)	**Awarded (2013-2015)** (n=28)	**Total** (n=179)
Official institution	66.1	68.8	69.5	66.9	75.0	68.2
Other, non-commercial organisation	33.9	53.1	44.1	45.7	35.7	44.1
Own source	23.2	14.1	28.8	21.2	25.0	21.8
Private company	14.3	18.8	22.0	17.9	21.4	18.4
Source not indicated	5.4	7.8	5.1	7.3	-	6.1

5 This is also much smaller a portion than the forty percent share Knight (2015: 65) found in data-driven stories from UK national newspapers.

6 Kruskal-Wallis test because of heteroscedasticity: $\chi 2 = 6.992$, $df = 2$, $p < .05$; pairwise Games-Howell tests revealed only one significant difference ($p < .10$) between 2013 and 2014.

As far as access to data is concerned, most of the analysed pieces that provide the respective information rely on data that is publicly available, just like in Parasie and Dagiral's (2013) study. This cannot be explained entirely by the fact that most data originate from official institutions, because with a share of 28.7 percent (*n* = 122), stories that draw on an official source are significantly more likely to report on data that had to be requested than stories not stating an official source at all (7.0%, *n* = 57, Fisher's exact test: $p < .001$; share of all cases: 21.8%, *n* = 179). Freedom of Information requests also belong in the category of requested data and were sometimes explicitly mentioned in the additional information about the data. Notwithstanding a drop in 2014, their number is rising again. So is that of the few cases based on data collected by the journalists themselves (see table 5). The share of leaked, requested and collected data is considerable. Yet, it does not appear as large as the link that scholars and practitioners often establish between data journalism and investigative reporting suggests (Parasie, 2015). Nonetheless, the portion of leaked information, as well as those of requested and self-generated data, are larger than those found by Knight (2015: 65) in her sample of data journalism in British national newspapers or by Tandoc and Oh (2015: 11) in their study of the *Guardian*'s *Datablog*. Furthermore, stories in our sample with requested or leaked information were significantly more likely to have a critical edge or a call for public intervention.[7] It is surprising that – despite data journalism's often cited association with openness and transparency (Coddington, 2015) – in over two-fifths of the pieces, journalists did not indicate at all how they accessed the data they used; in 2015 this was true for more than half of the analysed pieces.

Table 5: Access to data (multiple coding possible)

%	**2013** (n=56)	**2014** (n=64)	**2015** (n=9)	**Not awarded (2013-2015)** (n=151)	**Awarded (2013-2015)** (n=28)	**Total** (n=179)
Access to data not indicated	35.7	43.8	52.5	46.4	32.1	44.1
Publicly available data	39.3	43.8	40.7	41.7	39.3	41.3
Requested data	21.4	15.6	28.8	19.2	35.7	21.8
Own data collection	8.9[l]	1.6[l]	16.9[l]	7.3	17.9	8.9
Scraped data	5.4	7.8	5.1	6.6	3.6	6.1
Leaked data	1.8	4.7	3.4	2.6	7.1	3.4

[l] $\chi^2 = 8.929$; $df = 2$; $p < .05$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed only one significant difference between years 2014 and 2015 ($p < .01$).

7 Requested data: 87.2%, n = 39 vs. 38.6%, n = 140, Fisher's exact test: p < .001; leaked data: 100.0%, n = 6 vs. 47.4%, n = 173, Fisher's exact test: p < .05.

The data analysed in the stories refers to a range of geographical scales. Most notably, we found that while the share of projects drawing on international data has grown significantly over the years (2013: 10.7%, 2014: 15.6%, 2015: 32.2%),[8] the share of pieces based on regional information varies considerably (2013: 41.1%, 2014: 9.4%, 2015: 47.5%).[9]

In the majority of cases (88.3%), the data is analysed with a focus on comparing values (e.g. to show differences between men and women or neighbourhoods) and half of the pieces (50.3%) show changes over time (e.g. regarding global warming *Climate Change: How Hot Will It Get in My Lifetime?*). Connections (e.g. between a particular group of lawyers and the US supreme court) and flows (e.g. where Egyptian tax money went to) are illustrated in about a third of all projects (32.4%). Much less frequent (15.1%) are pieces that use data to show hierarchies – as in *Women as Academic Authors*, which ranks the most important female scientists. Although not statistically significant, the growth of the average number of different foci of analysis included in a story (2013: $M = 1.75$, $SD = 0.75$; 2014: $M = 1.80$, $SD = 0.76$; 2015: $M = 2.03$, $SD = 0.81$) indicates that data journalists increasingly combine these different approaches and perform more complex analyses.

Visualisation

If we think of data journalism as a distinct style of reporting, it is crucial to learn about the particular methods it uses to tell stories. Here, one of the most distinctive elements of data-driven pieces is the utilisation of visualisation techniques. These include tables and diagrams, such as pie charts or bar charts, that actually depict the data the story is based upon, but also non-data-related photos or illustrations that often serve nothing more than a decorative purpose. Table 6 shows that there is a more or less stable set of visualisation elements which mainly includes images and simple static charts (62.6% each) as well as maps (48.0%) and tables (33.5%); animated visualisations are rarer (16.8%). The proportion of images, charts and tables has grown significantly from 2013 to 2015. This partly echoes the findings of Appelgren and Nygren (2014), Parasie and Dagiral (2013) as well as Knight (2015: 65) that charts and maps are 'the most common form of data information presented'.

8 $\chi 2 = 9.412$, $df = 2$, $p < .01$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed only one significant difference between years 2013 and 2015 ($p < .01$).

9 $\chi 2 = 23.712$, $df = 2$, $p < .001$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed two significant differences between years 2013 and 2014 ($p < .001$) as well as between 2014 and 2015 ($p < .001$).

Table 6: Visualisation (multiple coding possible)[10]

%	**2013** (n=56)	**2014** (n=64)	**2015** (n=59)	**Not awarded** (2013-2015) (n=151)	**Awarded** (2013-2015) (n=28)	**Total** (n=179)
Image	46.4[I]	71.9[I]	67.8[I]	58.9[II]	82.1[II]	62.6
Simple static chart	55.4[III]	53.1[III]	79.7[III]	63.6	57.1	62.6
Map	51.8	46.9	45.8	49.7	39.3	48.0
Table	25.0[IV]	28.1[IV]	47.5[IV]	31.8	42.9	33.5
Combined static diagram	19.6	17.2	22.0	18.5	25.0	19.6
Animated visualisation	10.7	20.3	18.6	14.6	28.6	16.8
Other visualisation	-	-	8.5	3.3	-	2.8
No visualisation	-	-	1.7	0.7	-	0.6

[I] $\chi^2 = 9.284$; $df = 2$; $p < .01$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed two significant differences between years 2013 and 2014 ($p < .01$) as well as 2013 and 2015 ($p < .05$).
[II] Fisher's exact test: $p < .05$.
[III] $\chi^2 = 11.040$; $df = 2$; $p < .01$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed two significant differences between years 2013 and 2015 ($p < .01$) as well as 2014 and 2015 ($p < .01$).
[IV] $\chi^2 = 7.803$; $df = 2$; $p < .05$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed no significant differences.

On average, the pieces contained more than two different kinds of visualisations ($M = 2.46$, $SD = 1.12$). Moreover, this number has grown significantly over the years (2013: $M = 2.09$, $SD = 0.92$; 2014: $M = 2.38$, $SD = 1.13$; 2015: $M = 2.90$, $SD = 1.16$),[11] indicating that Knight's conclusion about data journalism in UK newspapers being 'practiced as much for its visual appeal as for its investigative qualities' (2015: 55) might also apply to this high-profile group of DJA-nominees. Typical combinations of visualising elements include simple static charts with images (39.7% of all cases) or maps (31.3%) as well as maps coupled with images (28.5%).

10 The numbers do not reflect whether elements of the same kind were included more than once: Several pictures, for instance, were counted as one visualisation of that kind.

11 $\chi 2 = 16.207$; $df = 2$; $p < .001$; Kruskal-Wallis test because of heteroscedasticity (Levene test). Games-Howell test revealed significant differences between: 2013 and 2015 ($p < .001$), 2014 and 2015 ($p < .05$).

Interactive features

Elements that allow users to interact with the data presented[12] are often discussed as another 'key characteristic' of data journalism (e.g. Coddington, 2015; Gray et al., 2012). However, our results are more in line with Tabary et al.'s (2016: 67) finding that 'data journalists focus on finding good quality data but engage very little with [...] interaction or reader participation' and often only 'integrate minimum formal interactivity': In our sample, 15.1 percent of cases offer no data-related interactive functions at all (see table 7). Yet, the average piece contains 1.67 different features (*SD* = 1.08), and only one of the award-winning projects provides no interactive feature at all. This leads us to speculate that interactivity is, nonetheless, considered a quality criterion.

Table 7: Interactive functions (multiple coding possible)

%	**2013** (n=56)	**2014** (n=64)	**2015** (n=59)	**Not awarded (2013-2015)** (n=151)	**Awarded (2013-2015)** (n=28)	**Total** (n=179)
No interactive functions	12.5	23.4	8.5	17.2[I]	3.6[I, II]	15.1
Zoom / details on demand	57.1[III]	54.7[III]	78.0[III]	62.9	64.3	63.1
Filtering	53.6	50.0	66.1	56.3	57.1	56.4
Search	30.4	23.4	27.1	28.5	17.9	26.8
Personalisation	23.2	14.1	15.3	15.2	28.6	17.3
Playful interaction	3.6	1.6	5.1	3.3	3.6	3.4

[I] Fisher's exact test: $p < .05$ (*one-sided*).
[II] One project: 'Reshaping New York'
[III] $\chi^2 = 8.401$; $df = 2$; $p < .05$; Fisher's exact tests for pairwise comparisons with adjusted α-levels (Bonferroni-Holm-correction) revealed only one significant difference between years 2014 and 2015 ($p < .01$).

The interactive features most often integrated into DDJ articles are zoom functions for maps, details on demand (e.g. the number of victims for each case of a reported school shooting), and filtering functions which allow the user to filter the provided data with respect to different variables (e.g. to only select voting results from one state or one year). Personalisation tools – where the user must enter personal data like their ZIP code or age to tailor the piece with customised data – are less common (17.3% of cases). Only six projects in the three years analysed include an opportunity for a

12 Features for follow-up communication, e.g. comment sections, that are often called interactive features, as well, fall into a different category ('opportunities for communication'; see table 2) and are not discussed in this paper.

gamified interaction (e.g. *Heart Saver*, a game in which the user must send ambulances as fast as possible to fictional characters having a heart attack).

Looking at developments over the years we find that, after a drop in 2014, the share of all interactive features has risen again in 2015, and on that basis, data journalism is becoming more interactive again.

Conclusion

In this paper we investigated if and how the emerging reporting style of data(-driven) journalism is developing using a content analysis of the pieces nominated for the *Data Journalism Awards* in the years 2013 to 2015. To advance our understanding of DDJ's evolution as a reporting style over the analysed time frame we identified the actors producing DDJ, the topics they cover and, in particular, the means they employ to do so, that is, the structural elements and forms of presentation.

The results on data journalism's development are ambivalent: On the one hand, our analysis shows that data journalism is both evolving and flexible in that different types of data, analyses and visualisation strategies are combined – or omitted – when it suits the topic and story. This echoes Coddington's observation that data journalists subordinate the use of data 'to the professional journalistic value of narrative and the "story"' (2015: 339). On the other hand, the set of potential elements to be combined appears to be stable and finite as we didn't find particularly striking developments and, over the years, we did not have to add new categories or variables to our initial codebook developed in 2013 to make sense of novel components. Instead, the new reporting style is (still) firmly characterised by those features that literature reviews, actor studies, and cursory observations have already hinted at. This finding might, of course, be induced in part by the method we used to produce it: quantitative content analyses are designed to reduce the complexity of their objects of investigation and are unable to detect developments that happen 'below the radar' of the variables and categories used. Here, further qualitative analyses could draw a more nuanced picture.

Another evidence for some degree of stability is that we found some 'typical combinations' reoccurring over the years. For instance, political stories are based significantly more often on polls and surveys than pieces on other subjects while business and economy topics are correlated with financial information, societal issues are covered using sociodemographic and geodata and health and science reports draw on measured values.

Above that, data journalism – at least in our sample – continues to be dominated by legacy print media and their online departments. The only other major players are investigative journalism organisations like *ProPublica* or the *ICIJ*. This finding, however, likely reflects the particular nature of our sample that – with its focus on award nominated pieces – is probably biased towards established and well-resourced actors in the field that are able to produce data-driven pieces to an award-worthy level – like the ones described by Parasie and Dagiral (2013), Karlsen and Stavelin (2014) and by Fink and Anderson (2015). When looking at how award-winning pieces differ from those that were only

nominated, we find that stories produced by investigative organisations and – although few in absolute numbers – by private or public broadcasters are represented above average among the award-winners. In contrast, projects by print magazines and news agencies, so far, have not been awarded at all.

Looking at the growing number of countries among the nominees of each year, it appears that data journalism is progressively spreading around the world. However, projects from the US and, to some extent from the UK, consistently make up the largest proportion of nominees; probably influenced by the fact that the DJA are issued by a global network of editors with English as their lingua franca and that the award website is in English, too. Moreover, we found that stories increasingly build on data gathered on an international scale and that they are often being published in two or more languages (one of them usually being English). On this basis, data journalism has the potential to foster the internationalisation of journalistic coverage and its distribution.

The average number of contributors to a data-driven piece has risen consistently while the share of projects involving external partners from outside the newsroom has fallen. This suggests to us that the production of data journalism, especially at the level of pieces nominated for an award, is progressively personnel intensive while the skills for it are increasingly being acquired within media organisations.

The DJA-nominees are characterised by an unchanged focus on political, societal, and economic issues. The small share of stories about education, culture and, especially, about sports – although in line with previous studies – might be unrepresentative of data journalism in general, but instead result from a bias towards 'serious' topics inherent in industry awards. In any case, neglected topics represent opportunities for expansion and innovation in the future, and, given the general trend of an increasing datafication of society, data journalism is likely to simultaneously extend coverage to more topics and domains based on ever more (publicly) available data sets.

Visualisations, the storytelling elements assumed to be most important in data-driven coverage, have maintained the same level of importance over the years and the average number of visual elements a piece comes with is still growing. However, there is an emphasis on rather simple types like maps and tables as well as on images with visual appeal but with little relation to the actual data. This opens up further avenues for innovation and distinction in the field.

The findings are similar for interactive features: Over the three years, they remain restricted to scalable maps, showing details on demand or filtering data by predetermined categories, while more sophisticated or gamified applications are rare.

In all of the three years under analysis, data journalism often assumed a watchdog role, containing elements of criticism or even calls for public intervention – and this stance of holding power to account and monitoring decisions and activities of politicians, corporations, and other socially important actors appears to be strengthened as we found

- a growing share of stories using more than one type of data (e.g. financial and sociodemographic data) or combining/contrasting data from different sources (e.g. official institutions and NGOs), and
- partial evidence that data journalists are looking more to other data sources besides official, openly accessible ones.

Our findings also illustrate how much data-driven journalism with a certain critical or watchdog attitude is appreciated by the DJA committee. Yet, these pieces are, more often than not, specifically based on publicly available data that does not even need to be investigated or 'uncovered' as such. Investigative approaches could be furthered by requesting data from institutions (e.g. through Freedom of Information requests) more often or collecting data oneself. This is especially true when considering that while data journalism spreads, there will likely be a branch of public relations that develops in parallel, or a culture of 'data-spin' that tries to influence coverage at the same time. For instance, reporter and data journalism educator Jonathan Stoneman (2015: n.p.) suggests that, '[w]hile Open Data is being touted by governments as their being open and transparent, journalists should be tracking what is really happening, what data are not being released, and why'. Moreover, the watchdog function of data journalism could be fostered even more by contrasting data from different social domains (e.g. contrasting numbers indicating worsening school education with the development of governmental spending on education) or by analysing data from their differing perspectives (e.g. looking at rising energy costs from both a business and an environmental perspective). Against this background, one can only imagine the potential that a branch of investigative and critical data journalism has to expose exploitation, corruption and the failures of power as some projects by *The Guardian, The New York Times,* and *ProPublica* among others have already demonstrated.

Overall, our findings suggest that data journalism as a reporting style with a certain set of core elements is, in fact, (still) evolving, but only partially, slowly and not in a linear way. We can assume that the analysed cases, as nominees for a DJA, fulfil a certain quality threshold and are considered examples of best practice within the field itself. As such, they are likely to co-determine the shape of data journalism to come (cf. Jenkins and Volz, 2016).

At this point, it is important to note that the question included in the title of this paper – 'Data-driven reporting – an on-going (r)evolution?' – can be interpreted in two different ways: Firstly, as outlined above, it asks to what extent data journalism is developing *internally*. Secondly, it might be understood as questioning the widespread notion that DDJ 'revolutionises' journalism in general by replacing traditional ways of discovering and reporting news. From this broader perspective, DDJ's development appears to be more of an evolution than a revolution: According to our findings, data-driven reporting is personnel-intensive and, by definition, reliant on the availability of data. As such, it cannot instantly react to breaking news. Additionally, due to its data dependency, it appears to neglect those social domains in which data are not regularly produced. Lacking such important characteristics of journalism – currentness and thematical universality – DDJ is more likely to complement traditional reporting than to replace it on a broad scale.

However, given the pace of innovation in the field, these observations are not much more than a snapshot. Moreover, while some of our findings could be interpreted as suggesting that data journalism is becoming more complex, we should bear in mind that the opposite is also true: The 'everyday' data-driven piece is increasingly easy to produce as more tools become available to help journalists get started.[13] Above that, DDJ's relevance and proliferation will certainly co-evolve with the increasing datafication of society as a whole.

13 See for for example, the 'Datawrapper': https://datawrapper.de/ (accessed 2 June 2016).

References

Anderson, C . W. (2013): Towards a sociology of computational and algorithmic journalism. *New Media & Society* 15(7): 1005–1021.

Appelgren, E. and Nygren, G. (2014) Data journalism in Sweden. Introducing new methods and genres of journalism into "old" organizations. *Digital Journalism* 2(3): 394–405.

Borges-Rey, E. (2016): Unravelling data journalism. A study of data journalism practice in British newsrooms. *Journalism Practice* 10(7): 833–843.

Boyles, J. L. and Meyer, E. (2016): Letting the data speak. Role perceptions of data journalists in fostering democratic conversation. *Digital Journalism* 4(7): 944–954.

Coddington, M. (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, J., Libert, M., Domingo, D., Heinderyckx, F. and Le Cam, F. (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.

Fink, K. and Anderson, C. W. (2015): Data journalism in the United States. Beyond the "usual suspects". *Journalism Studies* 6(4): 467–481.

Gray, J., Bounegru, L. and Chambers, L. (eds) (2012): The data journalism handbook. How journalists can use data to improve the news. (Early release). Sebastopol: O'Reilly.

Holtermann, H. (2011): *Datenjournalismus: eine neue Form der journalistischen Wertschöpfung aus Daten* [Data journalism: a new form of journalistically creating value from data]. Master Thesis, University of Hamburg, GER.

Jenkins, J. and Volz, Y. (2016): Players and contestation mechanisms in the journalism field. A historical analysis of journalism awards, 1960s to 2000s. *Journalism Studies*. Epub ahead of print 15 November 2016. DOI: 10.1080/1461670X.2016.1249008.

Karlsen, J. and Stavelin, E. (2014): Computational journalism in Norwegian newsrooms. *Journalism Practice* 8(1): 34–48.

Knight, M. (2015): Data journalism in the UK: a preliminary analysis of form and content. *Journal of Media Practice* 16(1): 55–72.

Krippendorff, K. (2013): *Content analysis: an introduction to its methodology.* Los Angeles: SAGE.

Lanosga, G. (2014): New views of investigative reporting in the twentieth century. *American Journalism* 31(4): 490–506.

Lewis, S. C. and Usher, N. (2014): Code, collaboration, and the future of journalism. A case study of the Hacks/Hackers global network. *Digital Journalism* 2(3): 383–393.

Parasie, S. (2015): Data-driven revelation? Epistemological tensions in investigative journalism in the age of "big data". *Digital Journalism* 3(3): 364–380.

Parasie, S. and Dagiral, E. (2013): Data-driven journalism and the public good. “Computer-assisted-reporters” and “programmer-journalists” in Chicago. *New Media & Society* 15(6): 853–871.

Royal, C. and Blasingame, D. (2015): Data journalism: an explication. *#ISOJ* 5(1): 24–46.

Stoneman, J. (2015): ‘Open data’ + journalism = ? *Data driven journalism. Where journalism meets data.* Available at: http://datadrivenjournalism.net/news_and_analysis/open_data_journalism#sthash.3s5QTVt6.dpuf (accessed 28 October 2015).

Tabary, C., Provost, A. M. and Trottier, A. (2016): Data journalism’s actors, practices and skills: A case study from Quebec. *Journalism: Theory, Practice, and Criticism* 17(1): 66–84.

Tandoc EC and Oh SK (2015): Small departures, big continuities? Norms, values, and routines in *The Guardian*’s big data journalism. *Journalism Studies.* Epub ahead of print 5 November 2015. DOI: 10.1080/1461670X.2015.1104260.

Wahl-Jorgensen, K. (2013): The strategic ritual of emotionality: a case study of Pulitzer Prize-winning articles. *Journalism: Theory, Practice, and Criticism* 14(1): 129–145.

Weinacht, S. and Spiller, R. (2014): Datenjournalismus in Deutschland. Eine explorative Untersuchung zu Rollenbildern von Datenjournalisten [Data-journalism in Germany. An exploratory study on the role conceptions of data-journalists]. *Publizistik* 59(4): 411–433.