Journal of Media Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rjmp20

Data journalism in the UK: a preliminary analysis of form and content

Megan Knight[a]
[a] School of Journalism and Media, University of Central Lancashire, Preston, Lancashire, PR2 2JQ, UK
Published online: 03 Mar 2015.

To cite this article: Megan Knight (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

To link to this article: http://dx.doi.org/10.1080/14682753.2015.1015801

Journal of Media Practice, 2015
Vol. 16, No. 1, 55–72, http://dx.doi.org/10.1080/14682753.2015.1015801

Data journalism in the UK: a preliminary analysis of form and content

Megan Knight*

School of Journalism and Media, University of Central Lancashire, Preston, Lancashire, PR2 2JQ, UK

More than two years ago Sir Tim Berners-Lee made the pronouncement that 'Journalists need to be data-savvy … but now it's also going to be about poring over data and equipping yourself with the tools to analyse it and picking out what's interesting'. 5This new form of data-driven journalism appears to have been enthusiastically adopted – at least in the rhetoric of news discourse, according to which it is 'rapidly becoming part of the establishment'. This analysis is a preliminary survey of data-based stories being presented in the national news in the UK, and lays the groundwork for an analysis and typology of the forms and formats of data journalism as a media practice. The analysis shows that while superficial data journalism is being practiced, it is limited in scope and format. No evidence was found of a commitment to data projects among the news outlets examined, and only one instance of recourse to the Freedom of Information Act was seen. Most data presented were superficial, and sourced from traditional outlets. Data journalism is practiced as much for its visual appeal as for its investigative qualities, and the overall impact, especially in the tabloid format is as much decorative as informative.

Introduction

Data journalism has become something of a buzzword in the last few years. Rather like citizen journalism and social media journalism before it, we have seen more and more reports that data journalism is the future, that journalists who cannot find and analyse complex data sets will find themselves the dinosaurs, left behind by this brave new world of media practice. As Sir Tim Berners-Lee said in 2010 'Data-driven journalism is the future' (Arthur 2010, para. 18). This claim, like others before it, needs examination. It is clear that there is more and more access to data for journalists: as the world becomes more digitised, more information is stored in data formats, and as freedom of information takes hold, at least in the developed world, more and more of that stored data will become accessible to the public and to journalists in one form or another.

History and development

'Data journalism' as a phrase seems to have appeared some time in 2008, in the *Guardian* newspaper. In December of that year Simon Rogers posted to the *Guardian Insider Blog* that:

*Email: maknight@uclan.ac.uk

> As of yesterday, our development team has come up with an application which takes the raw data and turns it into an editable map. Which meant that we could produce a fantastic interactive graphic based on these figures. It's data journalism – editorial and developers producing something technically interesting and that changes how we work and how we see data.(Rogers 2008, para. 7)

It is clear from this that although the specific software mentioned is new, the process of working with data is not a new one for Rogers, the *Guardian*, or the industry as a whole. Data Journalism appears to be the inheritor of two older news practices: infographics and computer-assisted reporting (CAR). News infographics – the production of graphs, charts, maps and other factual illustrations dates to at least the late nineteenth century. Simon Rogers claims that the *Guardian* staff have been doing data journalism since the newspaper's founding in 1821 (History of Data Journalism at the Guardian 2013). That table of information that had been leaked to the editor of the paper showed the costs of local schools, 70 years before compulsory education. Whether data journalism is something invented by the *Guardian* newspaper (as Rogers seems to be claiming, both in this video and elsewhere), it is probably uncontroversial to say that they are currently among the best known for doing data journalism. It is important to remember that the financial pages of newspapers have been publishing graphs, charts and tables of data for decades, and maps and other illustrations have been a feature of reporting for as long. *USA Today*, launched in 1982, revolutionised the newspaper graphic, bringing full colour, maps and boldness to the pages. *USA Today* has been criticised for dumbing down the news, reducing information to diagrams and pictures, but the use of graphics as an integral part of storytelling, not just in financial and weather reporting, did change the image of the newspaper and what it could look like (Barnhurst and Nerone 2001, 22; Friendly and Denis 2001; Gladney 1993).

CAR dates at least the 1980s, and the growth of personal computers, the Internet and of expertise in computing has contributed to this. In 1986, *Time Magazine* published a report on how 'in the computer age, newsmen are enlisting the machine with dramatic results' (Bowen 1986) which highlighted examples of computer analysis of data being used in investigative reports into riots, financial reports and fraud. In 1989 the National Institute for Computer-Assisted Reporting was founded in Missouri (Cox 2000). Philip Meyer is widely credited as one of the founders of CAR, and one of its earliest practitioners. In 1991 he published *The New Precision Journalism*, a book which located this new practice of computerised journalistic analysis firmly alongside the goals of objectivity, accuracy and the betterment of the journalistic profession (Meyer 1991). Although the book sets out a clear goal and meaning for this supposedly new form of journalism, its main goal is instructive: teaching students and journalists alike how to do it. This pattern remains true for much of the material published on CAR (and later on data journalism): the focus has been largely on the how, and far less on the why (or even the when). Meyer's book was followed by Brant Houston's (1999) *Computer-Assisted Reporting: A Practical Guide* and Matthew Reavy's (2001) *Introduction to Computer-Assisted Reporting*. These are textbooks, intended to provide instruction on how to do CAR. These books served to bring the idea of CAR into the mainstream of journalism, or at least of journalism education, and during the 1990s and early part of the first decade of the twenty-first century, considerable research was done into introducing CAR into

the journalism curriculum (Davenport, Fico, and DeFleur 2002; Lee and Fleming 1995; Miller 1998; Quinn 1997; Williams 1997) but less into the actual use of CAR in newsrooms.

A handful of researchers have looked at the use of CAR by working journalists, usually linking it explicitly to other technological changes: the adoption of computers in production and archiving (the digital morgue), and access to a wider range of electronic resources such as bulletin boards systems, the world wide web and government repositories of electronic information. Garrison's (2000a, 2000b) studies of the diffusion of electronic communication methods into newsrooms looked specifically at attitudes and training of staff, and the ways in which early adopters influenced the use of technology. An earlier study by him focused on the reasons why newsrooms had, or had not, adopted CAR: looking at organisation size (Garrison 1998) as a factor. Davenport, Fico, and Detwieler (2000) and Davenport, Fico, and Weinstock (1996) likewise looked at newsroom practices and structures, examining the prevalence of CAR in newsrooms with reference to circulation and infrastructure and followed it up four years later.

This line of research into the use of CAR appears to have been limited to the USA, and even to the areas of the USA in close proximity to the National Institute for Computer-Assisted Reporting in Missouri (and all published articles were produced for the American Association for Education in Journalism and Mass Communication or its journal). The studies, while useful, tend to be uncritical of the impact or desirability of CAR as a journalistic method, and focus instead on simple measurement of its use. What little discussion of the value of CAR is limited to unreferenced and unsubstantiated comments asserting its importance to the profession: 'Reporters using online databases and analyzing government data consistently won Pulitzers for their in-depth reporting' (Davenport, Fico, and Detwieler 2000, 3).

There are two articles that attempt to assess the value of CAR. Mayo and Leshner conducted an audience analysis of the credibility of newspapers using CAR, construction three versions of each a series of stories: one using CAR, one using anecdotal narrative and one using authoritative evidence. The subjects were then asked to rate the stories according to their credibility, newsworthiness, liking, quality and understanding. The readers did not rate the CAR stories any more or less credible or newsworthy than the others, but found them less likeable and readable, as well as being lower in quality. This study has not been repeated, but it raises interesting questions regarding the ways in which journalists perceive the impact or importance of a new technology to their profession, and the ways in which the audience perceive those changes (Mayo and Leshner 2000).

Maier's research into the use of mathematics in newspaper reporting is likewise a rather sobering read. In a study of the use of mathematical calculation in news stories he found that 48% of stories made mention of numerical information, and that fully one-third of those stories contained simple calculation errors, or 'errors of interpretation', including incongruence between charts and text, meaningless precision and 'naked numbers' (Maier 2010).

These two strands of the development of news production: increased use of graphics, and the availability of data and access to the means to analyse it continued through the first decade of the twenty-first century, but somewhat overshadowed by other technological developments in the field. In 2010 it was revived, though, apparently single-handedly by Simon Rogers and the team at the *Guardian*

newspaper, with help from Bradley Manning and Julian Assange. In July of that year, Wikileaks released the Afghan War Logs, followed by the Iraq War Logs in October to a number of news outlets, and then to the public. These massive data dumps contained hundreds of thousands of records of the activities of coalition troops in the two countries, and while damming, were frustratingly complex and detailed, and required a whole new level of analytical tools to make sense of them. The development of a custom data browser allowed the reporters to 'search stories for key words or events. Suddenly the dataset became accessible and generating stories became easier' (Rogers 2011, para. 261). The size of the data set was daunting, and making sense of it was a challenge – not just in terms of the management of the files, but also in terms of making the individual data points meaningful to readers. The *Guardian* used maps and charts to great effect with this data, and the apparently simple Iraq War Logs map of every death, made using Google Maps, remains one of the best examples of interactive data journalism around.

Rogers' book on the *Guardian*'s data journalism projects is one of the few published works on data journalism, or data-driven journalism. Like the others, it is written largely for practitioners, and remains somewhat uncritical of the impact of data journalism, or even aware of its actual use. Rogers is inevitably something of an evangelist for data journalism:

> So we are not alone in this: every day brings newer and more innovative journalists, developers and entrepreneurs into the field, and with them new skills and techniques. Not only is data journalism changing in itself, it's changing journalism too. And the world. (Rogers 2011, para. 40)

Other books on data journalism take a similar instructive line: focusing on the how, not the why or even the whether. *The Data Journalism Handbook* (Gray, Chambers, and Bounegru 2012) and Paul Bradshaw's (2010, 2011) work is likewise aimed at teaching people how to do it, and arguing for its inevitability in the newsrooms of the future. As with CAR, the focus is entirely on how to do it, and how amazingly revolutionary it is, but there is little critique of what data journalism actually is, who is actually doing it and why we should do it. This technological evangelism is not uncommon in journalistic research, and in journalism itself, but it needs more critical analysis.

This paper is a preliminary overview of the use of data journalism which will lay out the groundwork for a broader and more critical analysis of the prevalence, impact and value of data journalism as media practice.

A note on terminology

As with many other innovations in news production, there is considerable disagreement on what 'data journalism' actually is, or what the term encompasses. Data journalism and data-driven journalism are also routinely used as synonyms, while the older term, CAR has all but vanished (since it was coined at a time when 'computer' meant a mainframe beast occupying a whole room of the building, and now it is something we all have multiple examples of at our fingertips, this is unsurprising). Data journalism is defined by Simon Rogers (2011) as 'a field combining spreadsheets, graphics data analysis and the biggest news stories' (para.

110), while Mirko Lorenz (2010) refers to it as a process that goes from analysing, filtering and visualising data in a form that links to a narrative and is useful to the public. The emphasis on graphics and visualisation is common, and for some observers, data journalism is fundamentally the production of news graphics, and fits within that framework of practice, with elements of design and interactivity taking precedence (Bradshaw 2010; Lorenz 2010; Rogers 2011). For others, the focus on large data sources, often acquired through leaks or freedom of information requests, and the extended and complex analysis of this data is important, linking data journalism to the practice of investigative journalism, as Meyer (1991) did with CAR. For the purpose of this study, I have taken the broadest possible definition of data journalism: a story whose primary source or 'peg' is numeric (rather than anecdotal), or a story which contains a substantial element of data or visualisation. This broad definition allows for the widest possible net, catching as many examples as possible of journalism that incorporate data, in order to create an understanding of the field.

Methodology

The study is a content analysis of the use of data journalism in UK national daily and Sunday newspapers. The national papers were chosen because they are the best resourced, and prior work has shown that the size and resources of a newsroom are directly correlated to the extent to which those newsrooms make use of new technologies (Garrison 1998, 2000a; Machill and Beiler 2009; Quandt 2008). The newspapers are the *Guardian*, the *Times*, the *Daily Telegraph*, the *Independent*, the *Daily Mirror*, the *Express*, the *Sun*, the *Daily Mail*, the *Observer*, the *Sunday Times*, the *Sun on Sunday*, the *Sunday Telegraph*, the *Independent on Sunday*, the *Mail on Sunday* and the *Sunday Express*.

The print publications were used, because visualisations are not available in archive form, and online sites are either inaccessible to trawling software (as with News Corporation's publications), or contain little more than the print publication. The one exception to the latter is the *Guardian*, which is 'digital first', but although it remains in itself a fascinating study, the goal of the research is to examine the whole field, not its most extreme outlier. However, the print *Guardian* is included in the corpus, because to exclude it would skew the results. For the purposes of some of the analyses, the papers were combined into their respective ownership groupings, matching each daily with its sister Sunday title.

The newspapers were collected from 11 to 24 March inclusive, resulting in 112 newspapers. Each paper's main news section, lifestyle and entertainment sections were examined. The sports and finance/business sections were excluded from the study because their use and presentation of data are both historically much more entrenched, and because they follow substantially different processes and development of stories and narrative. Sports and financial stories that were covered within the main news section of the newspaper were included. Weather forecasts, and in the case of one newspaper, statistical analysis of the lottery numbers, were excluded on the grounds that their content is not journalistic, and the goal of this information is not the same as for journalism.

During the period of analysis the government's annual budget was announced (on 20th March), and coverage of the budget was included in the corpus, since it

always appeared within the main news section of the papers. Many of the papers had supplements for the budget, which were included, but specific coverage within the business sections were not.

The selection of these papers resulted in a corpus of more than 3000 stories, which were examined for the evidence of data journalism. Of the stories examined, only 106 had any element of data used within them. At this stage in the analysis, any story containing multiple elements or pieces, was counted as a single story: budget coverage was treated this way, which reduces the number considerably. The impetus for this first pass analysis was the origin of the story, or its peg, not the actual size of the content itself (Figure 1).

Three of the 'quality' newspaper groups, the *Guardian*, the *Times* and the *Independent*, account for 68 of the stories, 64% of the total, with the last of the qualities, the *Telegraph*, making up another 9 stories. The 'popular' papers had a far lower number of data-driven stories, reasonably evenly split among the members of that group. On average, the quality papers had slightly fewer than one data-driven

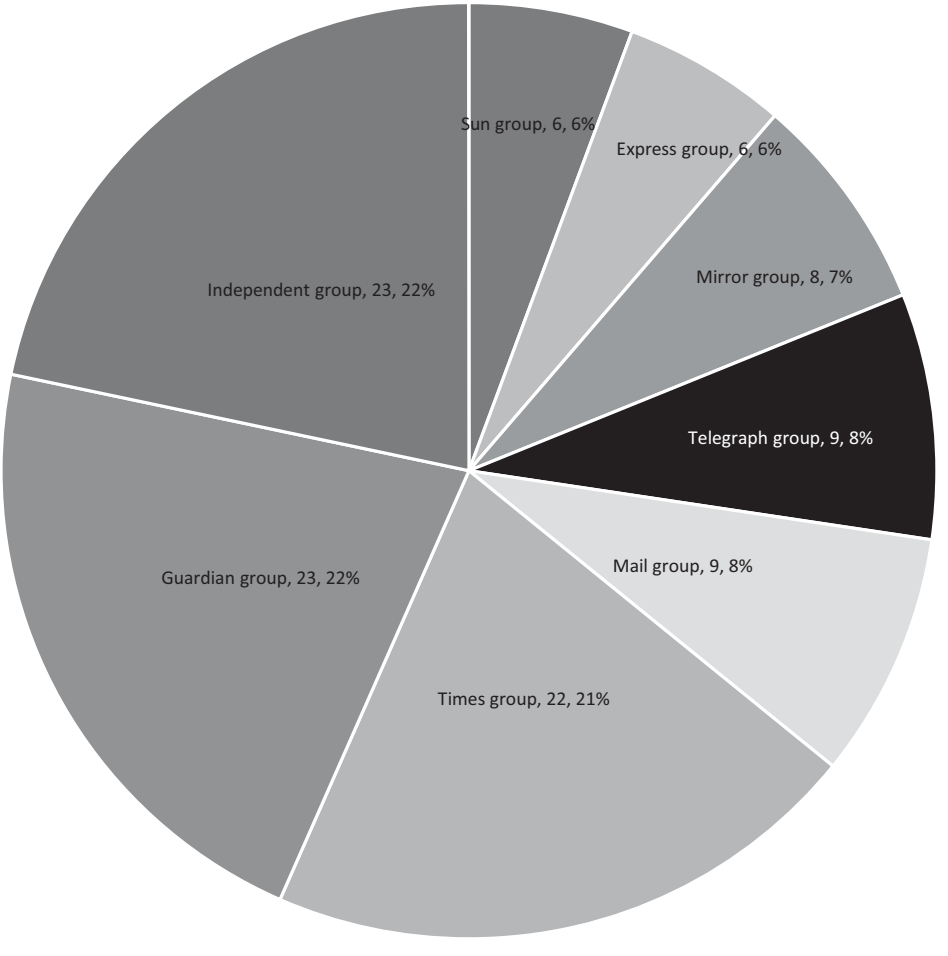

Figure 1. Data-driven stories in all publications, grouped by ownership.

story per day, but that was not evenly split across the two weeks. Leaving aside the budget coverage (which all appeared on the 21st of March), the breakdown of stories across the weeks was uneven, with more stories appearing on Friday and Sunday than other days.

Findings

Stories were categorised by their main subject (not into their identified section within the newspaper). Stories covering social issues (poverty, the environment, education and housing) were disproportionately likely to contain data elements (for reasons discussed below), followed by world and news stories. There were relatively few science stories, but data elements represent a substantial proportion of the science stories covered (Figure 2).

Each story was then analysed as to the data components, or elements that each contained. Many stories contained multiple data elements, so this analysis allowed a clearer idea of the extent of data reporting within the corpus. There were 172 data elements presented within the 106 stories, but once the budget stories were excluded, there were 111 data elements in 98 stories. Examination of the individual data elements resulted in the development of the following categories, or types, of data element: a textual analysis, where the numbers are discussed within the text, but not otherwise represented; timeline, which shows events listed by date, whether continuous or not; static map, showing the location of an event; dynamic map, showing both location and other data such as amount or date; graph, showing relationships between numbers (these were initially broken down by graph type, but this resulted in too fine a division of data); infographic, a combination of pictures and numerical information; a table of figures, a list of numbers and a numerical pullquote (Figure 3).[1]

Overall, infographics, graphs, chart, static maps and pullquotes were the most common form of data information presented, with little variation among them. The less accessible, and graphically interesting, forms of data, lists, tables and textual analysis were less common. There is a strong prevalence for the visual impact in the

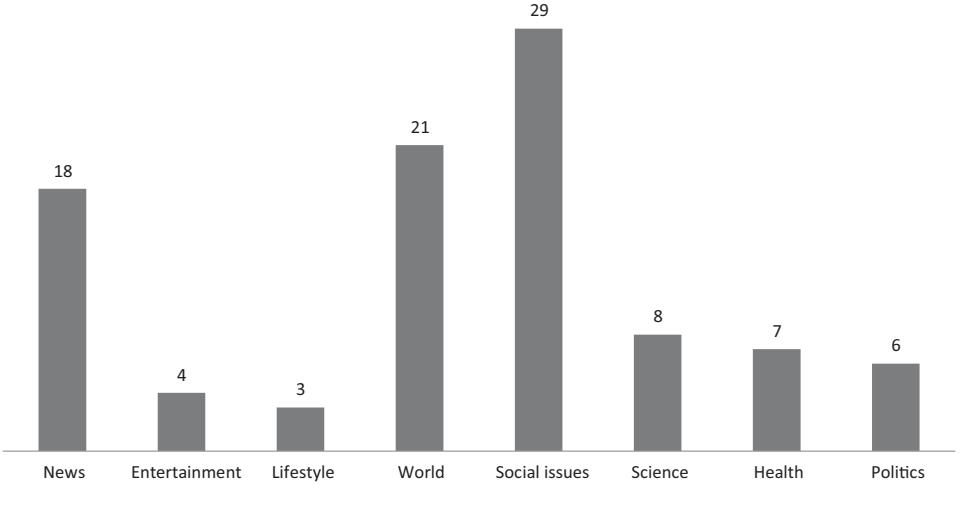

Figure 2. Subjects covered by data-based stories.

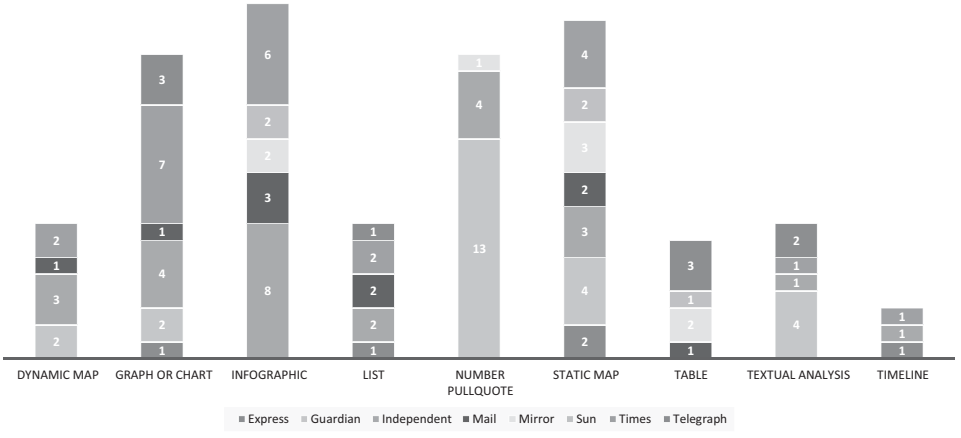

Figure 3. Data-driven types in all publications, grouped by ownership.

data elements presented: especially among the popular titles. The least visual elements, the number pullquote and textual analysis appear only in the quality papers. The number pullquote is almost entirely the preserve of the two left-of-centre quality titles: the *Guardian* and the *Independent*, with only one appearing in the *Mail*.

Static maps were the most evenly spread across the titles – all outlets had at least one map element, although the more complex dynamic maps only appeared in the *Guardian*, the *Independent*, the *Times* and the *Mail* (Figure 4).

Infographics were used in all subjects, except health. As expected, maps were used most in world stories, followed by news (and then by science – but there were four stories on the building of a new telescope array in Chile, all contained maps of the location, which is a somewhat anomalous usage); pullquotes (a quote or piece of information from the story displayed as a visual element on the page – in this case this refers to a pullquote containing key numeric information) were commonly used in health reporting, but given that pullquotes were used only by two titles, the *Guardian* and the *Independent*, and that health stories were disproportionately covered by those two titles, it is appears that this is not a function of the nature of the stories. The remainder of the types was fairly evenly distributed among the subject matter.

Budget coverage

The budget coverage across all eight titles contained more graphs and charts than any other type of element, as would be expected, as well as half of the total tables presented in the corpus. As with other subjects, number pullquotes are extensively used, primarily by the *Guardian* and the *Telegraph* (the only time the *Telegraph* used pullquotes was in budget coverage). As expected maps and lists were hardly used, and timelines not at all (Figure 5).

The quality papers had far more extensive and far more numeric budget coverage than the popular titles. All of the quality titles had budget supplements as well as coverage in the main body: total coverage in each of the quality titles averaged 23 pages (the *Telegraph*, being the last surviving broadsheet daily, was calculated at the equivalent of two tabloid pages per page). The popular titles averaged 7.5 pages of

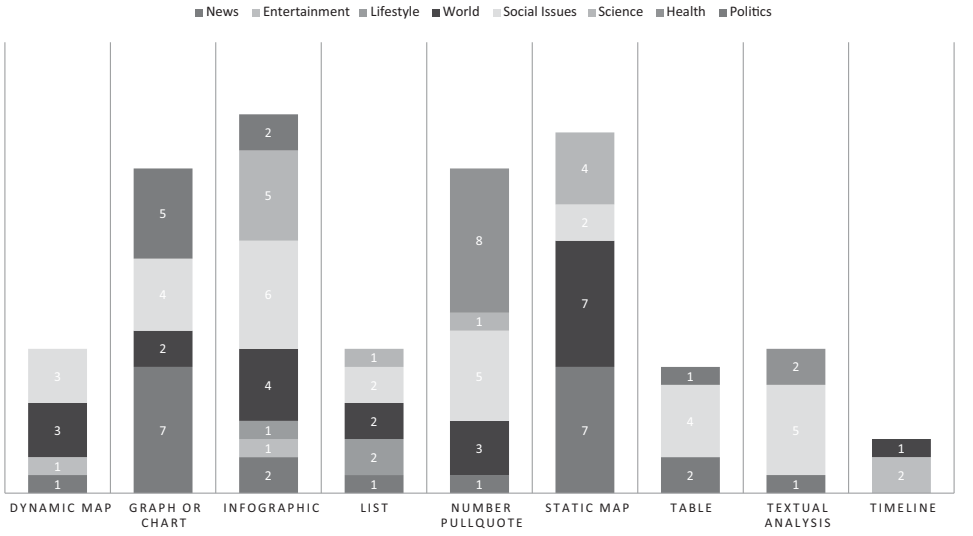

Figure 4. Types of data elements used in each subject area.

coverage, and none them included separate supplements. A ratio of data elements per page of coverage was calculated, with the following results (Figure 6).

The *Guardian* and the *Telegraph* had the highest ratios, even with number pullquotes removed from the analysis, the *Guardian* had a ratio of 0.6 elements per page of coverage. The relatively low ratio within the overall corpus is somewhat surprising, given the numeric nature of the subject. Most of the coverage, however, was given to narrative discussion of the impact of the budget on the public, on business and on the various political parties. All supplements did contain large graphic elements, but the actual data within them were somewhat limited (most supplements contained some variation on a pie chart covering two pages, showing income and expenditure: dramatic, but that single element contained only 25 or so data points within it).

Sources of data

The sourcing of data is widely considered to be a key part of data journalism, and certainly the best known data journalism investigations are remarkable for the nature of the data, and how the news organisations came by it. The increasing power of freedom of information acts in the developed world has resulted in more data being released to news organisations in that way, this, along with the now reasonably common dumps of leaked data have led to something of a perception that data journalism is all about massive data sets, acquired through acts of journalistic bravery and derring-do (Bradshaw 2010; Leigh and Harding 2011). Data journalism is also explicitly linked to investigative journalism (and has been since Meyer), and to the importance of quality, original journalism and journalism which seeks to 'speak truth to power'. At the very least, 'data journalism is all about diverse sources', according to Simon Rogers (2011, para. 53).

The origin of the data used in each data element within the corpus was examined. Several generic data sources were identified, including government, corporate entities, research institutes (including non-profit groups and academic institutions),

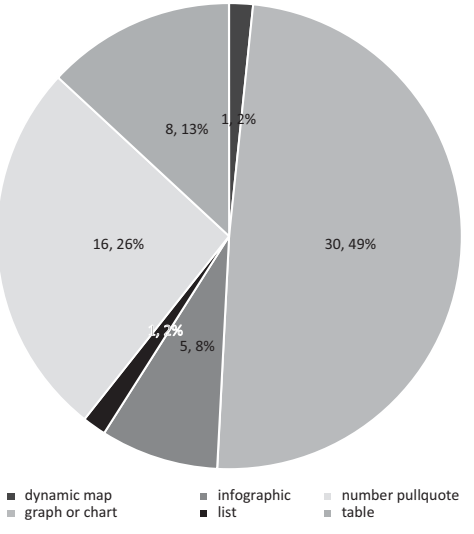

Figure 5. Types of data elements used in budget coverage.

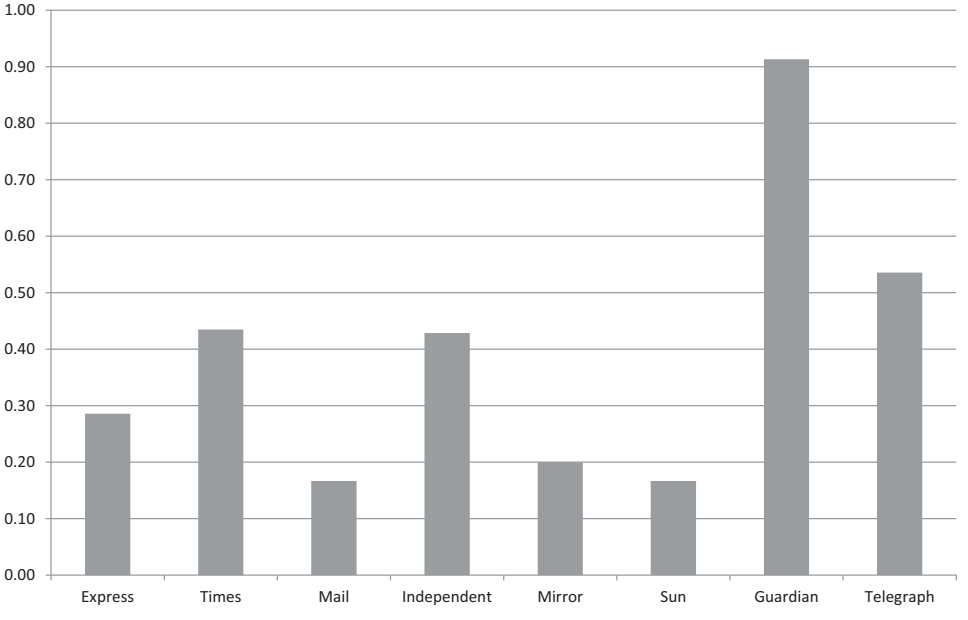

Figure 6. Ratio of data elements per page of coverage.

Pan-national organisations (such as the various agencies of the United Nations), Polls and self-generated (i.e., data gathered by the news organisation itself). This is of necessity a rough measure, but it fits in line with examining the claim that data journalism is a way to break away from the dominance of official sources and press releases.

Since all the data presented in the budget coverage were sourced via government, this was excluded from the analysis, of the remaining elements, 40% were unsourced, 28% from a research institution, 11% from a government department, 7% each from polls and self-generated by the organisation, and only two tranches of data, from more than 100, were gained through either a freedom of information request or a leak.

Of the data elements for which no source was identified, almost half (19 of 41) were static maps, which would require no specific source be identified. A further nine were infographics, and the remainder a range of data. In some cases, the data were uncontroversial (such as the size of various atomic elements), and thus required no specific source, or clearly sourced via another news organisation (the decline of the convention that all wire service stories be clearly identified has made the work of academic research considerably harder).

The corpus showed a heavy reliance on data sourced through research institutions, especially in stories covering health and social issues. All of the health stories and a third of those covering social issues sourced their data through research institutions, and in half of those instances, the data were clearly included in a press release issued by the research institution and represented without comment or challenge by the journalist.

The self-generated data comprised information collated from consumer products (prices, calorie counts, etc.) or data acquired during traditional interview techniques

(how a celebrity spends their money, for example). The one large story that relied on self-generated data was an 'investigation' in the *Sun* on 11 March, called 'Psychic Britain', spread across two pages, with a large infographic covering more than a page of area. The investigation was in fact a poll of *Sun* readers conducted by the newspaper (no polling organisation is identified). The text discusses psychics, giving a brief history, and profiling four psychic practitioners. No numeric data are included in the text. The infographic shows the results of 11 questions, 3 shown as pie charts and 8 yes/no questions such as 'Do you believe there are buildings which are haunted?' No historical or comparative data are given, and no details of how the poll was conducted are provided.

The two stories based on freedom of information requests appeared in tabloid newspapers.

A feature in the *Mirror* on 22 March, headlined 'Teen Sex Plague' covered two full pages, with an infographic showing rates of diagnosis of sexually transmitted diseases in under 16s and under 13s, in some cases split by gender, in 2003, 2007 and 2011, alongside a table of rates of specific diseases, and three key figures pulled out as subheadings. The design of the infographic has clearly sacrificed clarity for impact – the key figures are rendered as pie charts, superimposed on condoms and in a bar chart rendered as test tubes. Although the clarity and effectiveness of data visualisations are not the intent of this article, it must be pointed out that using a pie chart to show changes in figures over time is not at all effective, and renders the numbers essentially meaningless. The text accompanying the feature contains approximately one-quarter numerical data, but at the end of main story, and not at all in either of the two accompanying pieces.

The other large original story was in the *Mail on Sunday* on 17 March, headlined: 'The Great Green Con', and concerned figures 'leaked' by the Intergovernmental Panel on Climate Change (IPCC), although the paper does not make the route clear, and other sources refer to the data as having been released. The only data discussed in the story are in the form of a complex graph, taking up more than a third of the two-page spread, and showing predicted global temperatures according to the IPCC, mapped to actual recorded temperatures. The point at which the recorded temperatures appear to dip below predicted temperatures is highlighted. The text contains almost no data itself, focusing on the politics of the claims. The story is big, and is given a lot of space, but it must be said that the newspaper's interpretation of the data has been heavily criticised in other media, and by several reputable organisations.

Analysis

A simple content analysis of the findings above shows a variety of data elements used, in a variety of subjects. The more subtle questions of the value added by the use of data journalism require a more complex understanding of two aspects of data journalism: interpretation and visualisation.

Interpretation of the data is often identified as a requirement of data journalism, although this is not uniformly supported within the field. Bradshaw (2010) and Rogers (2011), certainly, identify the interpretation of complex data sets as one of the skills of a data journalist, as does Lorenz (2010), but they also both suggest that the presentation of raw data sets is also data journalism.

Complexity

Ranking the data element types by the level of interpretation and analysis required to produce them (based on a reading of the text, of the amount of information included in the element, and the proportion of the story given over to the data itself), we get the following hierarchy of value added to the story by the process of data journalism, from least to most complex, with a numerical value attached:

(1) Number pullquote – a single numerical fact, presented out of context and without comment
(1) Static map – a location identifier, a graphical dateline, with one or more locations identified
(2) List and timelines – a one dimensional ranking of a series of data points
(3) Table – a two-dimensional presentation of data in a grid format. This is arguably more complex than a graph or chart, but it requires less analysis or interpretation on the part of the journalist.
(3) and (4) Graphs and charts – a visual representation of two-dimensional information. These were further divided into simple, and complex data sets.
(4) Dynamic map – a map showing locations in relationship to time or other values
(5) Textual analysis – a complex discussion of numerical information in the text
(5) Infographic

This ranking allows for a more nuanced understanding of the complexity of data journalism present in the corpus, and analysing the elements by this measure, gives us mean complexity scores for each newspaper title (Figure 7).

Although the *Telegraph* and the *Mail* had relatively few data elements within the corpus, those they had were more complex and nuanced. The *Guardian*'s reliance on number pullquotes has brought its complexity score down considerably, without them included in the analysis, its mean complexity score is among the highest, at 3.6.

The *Mail* and the *Sun* newspapers score surprisingly high on complexity, this is accounted for by their use of infographics.

Visual appeal

Data journalism is inextricably bound up with the visualisation of data, and in the corpus it is clear that some data choices have been made to increase the visual appeal of the material, at the expense of clarity of data. Newspapers are both visual and textual, and the importance of design and images to the newspaper industry should not be minimised, but some data elements were clearly designed to be primarily eye-catching, with little concern for the intelligibility of the finished product. The types of data element were ranked according to their visual appeal, in the following hierarchy:

(0) Textual analysis
(1) Number pullquote or table
(2) Timeline or list (although a list is arguably less visual than a table, examination of the elements shows that lists were almost always combined with images)
(3) Static map, chart or graph
(4) Dynamic map
(5) Infographic

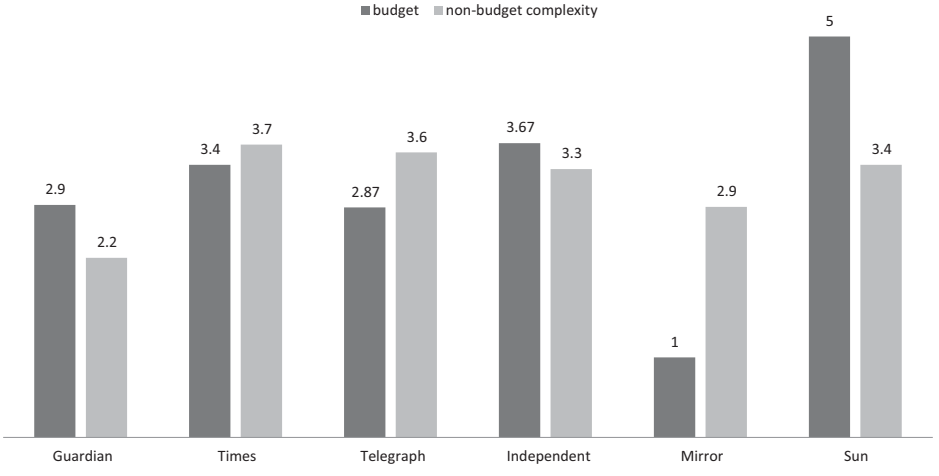

Figure 7. Complexity rating for each news outlet's data journalism.

A mean visual appeal score is calculated for all titles, budget and non-budget coverage, with results as follows (Figure 8).

As with the mean complexity score, the *Guardian*'s result is somewhat skewed by it heavy reliance on number pullquotes: removing those, however, only raises its mean visual appeal to 1.7, still the lowest of the newspapers. The *Telegraph* is in second place, but the remaining two quality papers, the *Independent* and the *Times* are ranked comparatively with the popular titles, with a mean visual appeal of between two and three. The *Times*' reliance on infographics accounts at least in part for this.

For the budget coverage all papers displayed slightly less visual appeal than for non-budget stories, with the exception of the *Sun*, again, as a result of their use of infographics (the only data element in their budget coverage was a large infographic). When calculated against the overall ratio of data journalism to coverage, the

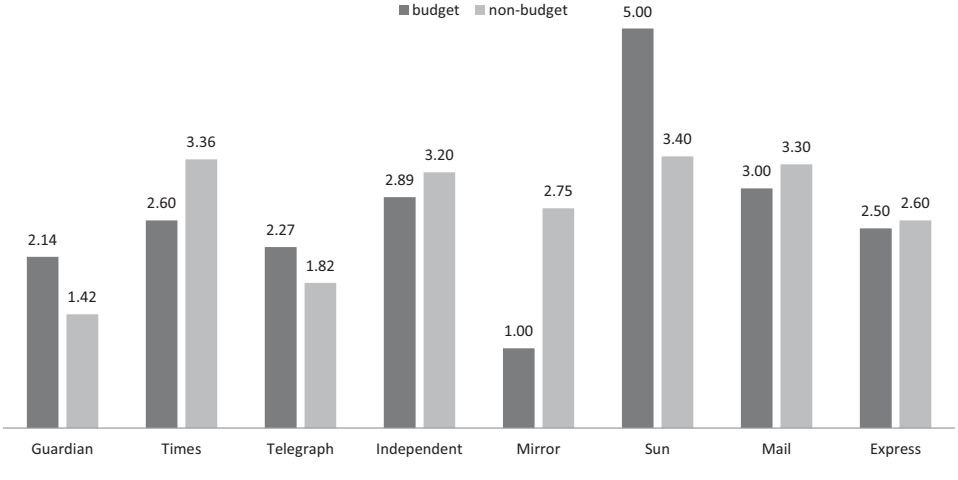

Figure 8. Visual appeal rating for each news outlet's data journalism.

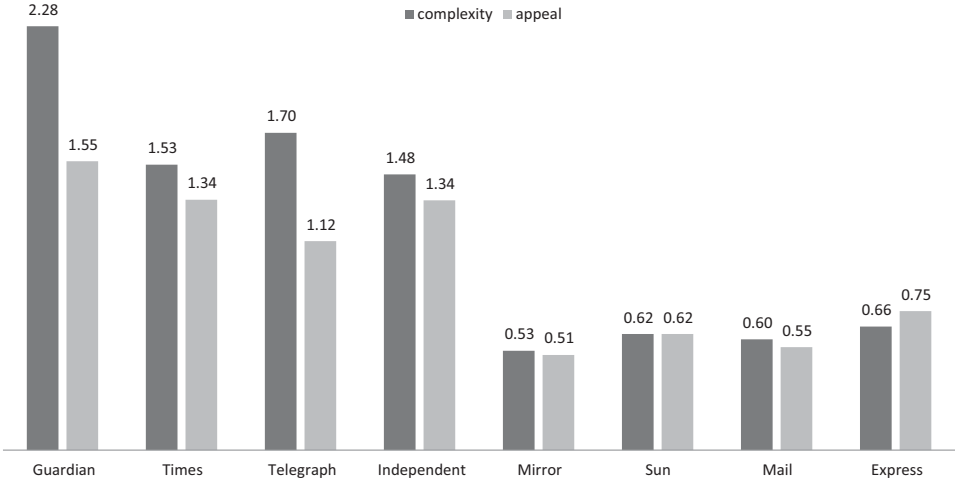

Figure 9. Visual appeal and complexity of the data journalism presented, as in index of the ratio of data journalism.

complexity and visual appeals scores are reduced. This final calculation, of the complexity of the data journalism presented, as a function of the proportion of data journalism overall, shows the following (Figure 9).

Overall, the corpus displays only slight differences in either complexity or visual appeal between the various titles, or between budget and non-budget coverage. The substantial difference is in the use of data journalism at all, which is far more evident in the quality papers' coverage than that of the popular press.

Conclusions

The study is examines and highlights some of the claims made by data journalism evangelists against the reality. The aim is not to demolish the claims of data journalism, but to examine its level of penetration into (mainstream news) media practice. To that end, the findings do not show overwhelming evidence of comprehensive use of data journalism by national UK titles, although all titles did make some use of data journalism throughout the period studied. As expected, the *Guardian* newspaper shows more commitment to data journalism, and to more complex data journalism than other titles, with the remaining quality titles in similar rankings. The popular newspapers show a lower commitment overall to data journalism, but appear to value visual appeal and complexity equally.

The data journalism presented relied heavily on institutional sources, especially government agencies. In addition, there is evidence of the rise of data-based press releases: a substantial proportion of the stories showed evidence of a body of data being released wholesale by research institutes and other vested bodies. Particularly in social issues and health, most of the data presented were acquired in this way.

Freedom of Information actions and leaks were not widely represented in the data presented, although there was some small evidence. Large dumps of data acquired through investigative journalism are rare, and a much longer study is needed in order to capture them. What is more concerning, though, is the appropriation of the

language of investigative data journalism for somewhat less rigorous material. The *Mail on Sunday*'s controversial analysis of IPCC data it claims had been leaked to it, and the *Sun*'s presentation of a poll that amounts to asking its readers whether they believe in ghosts were both presented as 'special investigations' purporting to uncover hidden evidence. This shows that although the methods espoused by the evangelists of data journalism are not being widely followed, the form has its own appeal, and that the presentation of information in data form can have its own weight, regardless of the actual value of the information or its impact on society. The extent to which commercial newspapers actually engage with investigative journalism, as opposed to the extent to which they claim they do is a subject for another study, but an important consideration in the analysis of news reporting.

Overall, though, the data journalism found in this study is largely superficial, institutionally sourced and non-remarkable. Rather than becoming the new frontier of investigative journalism, this very limited study has shown that in the daily newsroom grind, crunching data have become no more remarkable, or important, than any other form of journalism.

Disclosure statement

No potential conflict of interest was reported by the author.

Note

1. A pullquote is an excerpt from the body of the text repeated and highlighted on the page to draw the eye. A numerical pullquote is extract of one piece of data treated in this manner.

Notes on contributor

Megan Knight is a Senior Lecturer in International Journalism at the University of Central Lancashire in Preston, UK. She has worked extensively as a journalist and as a webmaster for news organisations ranging from alternative weeklies in Vancouver to the daily *Star*, the *Sunday Independent* and the South African Broadcasting Corporation in Johannesburg. The former director of the New Media Lab at Rhodes University, and of the Highway Africa Conference, she has research interests in new media technologies and especially their effects on the professional identity of journalists, as well as in alternative and radical media. She is the co-author (with Clare Cook) of Social Media for Journalists: Principles and Practice, London: Sage, 2013.

References

Arthur, C. 2010. "Analysing Data Is the Future for Journalists, Says Tim Berners-Lee." *The Guardian Online*, November 22. Accessed January 28, 2013. http://www.Guardian.co.uk/media/2010/nov/22/data-analysis-tim-berners-lee.

Barnhurst, K. G., and J. C. Nerone. 2001. *The Form of News: A History*. New York: Guilford Press.

Bowen, E. 1986. "Press: New Paths to Buried Treasure." *Time Magazine*, July 7.

Bradshaw, P. 2010. "How to Be a Data Journalist | News | Guardian.co.uk." *The Guardian Data Journalism*, October 1. Accessed November 30, 2012. http://www.Guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide.

Bradshaw, P. 2011. *The Online Journalism Handbook: Skills to Survive and Thrive in the Digital Age*. 1st ed. Harlow: Longman.

Cox, M. 2000. "The Development of Computer-assisted Reporting." Paper presented at the Association for Education in Journalism and Mass Communication, Southeast Colloquium, Chapel Hill, March. http://com.miami.edu/car/cox00.pdf.
Davenport, L., F. Fico, and M. DeFleur. 2002. "Computer-assisted Reporting in Classrooms: A Decade of Diffusion and a Comparison to Newsrooms." *Journalism and Mass Communication Educator* 57 (1): 6–22. doi:10.1177/107769580205700103.
Davenport, L., F. Fico, and M. Detwieler. 2000. "Computer-assisted Reporting in Michigan Daily Newspapers: More than a Decade of Adoption." Paper presented at the AEJMC National Convention, Phoenix, Arizona, August. http://online.sfsu.edu/jjohnson/Courses&-Syllabi/BU-JO807/Bibli&Articles/Davenport(2000).pdf.
Davenport, L., F. Fico, and D. Weinstock. 1996. "Computers in Newsrooms of Michigan's Newspapers." *Newspaper Research Journal* 17 (3): 14–28.
Friendly, M., and D. J. Denis. 2001. *Milestones in the History of Thematic Cartography, Statistical Graphics, and Data Visualization*. Toronto: York University. http://euclid.psych.yorku.ca/SCS/Gallery/milestone/.
Garrison, B. 2000a. "Diffusion of a New Technology: On-line Research in Newspaper Newsrooms." *Convergence: The International Journal of Research into New Media Technologies* 6 (1): 84–105. doi:10.1177/135485650000600109.
Garrison, B. 2000b. "Journalists' Perceptions of Online Information-gathering Problems." *Journalism & Mass Communication Quarterly* 77 (3): 500–514. doi:10.1177/107769900007700303.
Garrison, B. 1998. "Newspaper Size as a Factor in Use of Computer-assisted Reporting." Paper presented at the Association for Education in Journalism and Mass Communication, Baltimore, August, 1998.
Gladney, G. A. 1993. USA Today, Its Imitators, and Its Critics: Do Newsroom Staffs Face an Ethical Dilemma? *Journal of Mass Media Ethics* 8 (1): 17–36. doi:10.1207/s15327728jmme0801_2.
Gray, J., L. Chambers, and L. Bounegru. 2012. *The Data Journalism Handbook*. Sebastopol, CA: O'Reilly Media.
History of data journalism at the Guardian. 2013. London. http://www.Guardian.co.uk/news/datablog/video/2013/apr/04/history-of-data-journalism-video.
Houston, B. 1999. *Computer-assisted Reporting: A Practical Guide*. 2nd ed. Boston: Bedford/St. Martin's.
Lee, K. C., and C. A. Fleming. 1995. "Problems of Introducing Courses in Computer-assisted Reporting." *Journalism and Mass Communication Educator* 50 (3): 23–34. doi:10.1177/107769589505000304.
Leigh, D., & L. Harding. 2011. *Wikileaks: Inside Julian Assange's War on Secrecy*. 1st ed. New York: Public Affairs.
Lorenz, M. 2010. "Data Driven Journalism: What Is There to Learn?" Paper presented at the IJ-7 Innovation Journalism Conference, Stanford, CA, June.
Machill, M., and M. Beiler. 2009. "The Importance of the Internet for Journalistic Research." *Journalism Studies* 10 (2): 178–203. doi:10.1080/14616700802337768
Maier, S. 2010. "All the News Fit to Post? Comparing News Content on the Web to Newspapers, Television, and Radio." *Journalism & Mass Communication Quarterly* 87 (3/4): 548–562. doi:10.1177/107769901008700307.
Mayo, J., and G. Leshner. 2000. Assessing the Credibility of Computer-assisted Reporting." *Newspaper Research Journal* 21 (4): 68–82.
Meyer, P. 1991. *Precision Journalism: A Reporter's Introduction to Social Science Methods*. Lanham, MD: Rowman & Littlefield.
Miller, L. C. 1998. *Power Journalism: Computer-assisted Reporting*. Fort Worth, TX: Harcourt Brace College.
Quandt, T. 2008. "(No) News on the World Wide Web." *Journalism Studies* 9 (5): 717–738. doi:10.1080/14616700802207664.
Quinn, S. 1997. "Learning the 4Rs of Computer-assisted Reporting in Australia." *Asia Pacific Media Educator* 1 (3): 131–141.
Reavy, M. 2001. *Introduction to Computer-assisted Reporting: A Journalist's Guide*. Mountain View, CA: Mayfield.

Rogers, S. 2008. "Turning Official Figures into Understandable Graphics, at the Press of a Button." *Inside the Guardian Blog. Newspaper*, December 18. http://www.Guardian.co.uk/help/insideGuardian/2008/dec/18/unemploymentdata.

Rogers, S. 2011. *Facts Are Sacred: The Power of Data*. London: Guardian Books.

Williams, W. S. 1997. "Computer-assisted Reporting and the Journalism Curriculum." *Journalism and Mass Communication Educator* 52 (1): 67–71. doi:10.1177/107769589705200108.