Reporters in the age of data journalism

Andreas Veglis and Charalampos Bratsas, Aristotle University of Thessaloniki

Abstract

In the past, journalists were responsible for reporting the news. But today news stories disseminate as the incidents unfold, from multiple sources. Thus, gathering, filtering and visualizing events has a growing value. Huge amounts of data are available, but exploiting them is not an easy task. Data journalism can be defined as a journalism speciality in which numerical data are used in the production and distribution of information. This article investigates the necessary skills that journalists must have in order to cope with data journalism. More precisely, it defines data journalism, and discusses journalists’ Information and Communication Technology (ICT) skills, as well as the necessary skills for supporting data journalism. Special attention is given to Web 3.0 and open data that can play an important role in data journalism. A survey conducted among professional journalists in Greece concerning data journalism is also presented and discussed.

Keywords

data journalism
Introduction

The introduction of Information and Communication Technologies (ICTs) has affected literally every aspect of human activity. Thus, journalism has been transformed by digitalization of the work process as well as by the introduction of the Internet, along with its services (Veglis 2009). The evolution of citizen journalism has increased competition, and traditional media companies are under pressure to deliver more and faster news (Allan and Thorsen 2009). Today, the journalist is expected to have the ability to: exploit many tools and services in order to be instantly informed about breaking news as well as current events; and use a variety of tools and applications to prepare and disseminate news articles. New types of journalism have emerged – namely multimedia journalism (Bull 2010) and data journalism (Gray et al. 2012) – which require journalists to have special ICT skills.

Journalism has always been considered to be closely related to technology. Radio, television and in the last decade of the twentieth century, ICTs have expanded their influence and also created new opportunities for gathering and consuming, as well as for creating and disseminating, news (Spyridou et al. 2013).
In recent years data journalism has garnered significant attention both in the academic literature and in the area of new developments in digital news production (Appelgrena and Nygren 2014; Fink and Anderson 2014; Mair and Keeble 2014). As in the case of social media journalism and citizen journalism, data journalism is widely considered the future of journalism (Knight 2015). Data journalism is a new form of journalism that has evolved gradually over the past years, driven by the availability of data in digital form. In today’s digital world almost everything can be described with numbers (Gray et al. 2012). Data journalism is a journalism speciality reflecting the increased role of numerical data in the production and distribution of information in the digital era (Thibodeaux 2011). Data can be the source of data journalism or/and it can be the tool with which the story is told (Gray et al. 2012).

This article studies the issue of data journalism. First, the state of data journalism in Greece is examined with the help of a survey that focuses on data journalism practices among professional journalists. In subsequent sections data journalism is defined and the necessary ICT skills that journalists should have in order to cope with the needs of data journalism are discussed. These skills are closely associated with the stages in the development of a data journalism project. A thorough analysis of Web 3.0 skills that journalists may possess is also included. Conclusions and future extension of this work are presented in the last section.

A survey for the use of data journalism in Greece

Data journalism in Greece

Data journalism initially appeared in Greece in the form of simple infographics that informed the audience about certain parameters of news stories. The majority of these
visualizations are static and they are published in print editions and on the web. We have very few examples of middle-scale data journalism projects that disseminate substantial amounts of information. Some interesting projects were produced in post-election periods and included analytic election results (e.g., http://www.skai.gr/ekloges2012, http://www.igraphics.gr/infographics/greece). The concept of data journalism is not widespread in Greece. The term Data Journalism is rarely mentioned in Greek media. In 2014 Veglis and Bratsas translated the Data Journalism Handbook (J. Gray et al., 2012) by the European Journalism Centre (EJC) into Greek language.\textsuperscript{1} Further, in the spring of 2014 an elective course on Data Journalism was taught in the School of Journalism and Mass Communication at Aristotle University of Thessaloniki.

**Methodology**

In order to fully investigate the state of data journalism in Greece a survey was conducted with the help of an online questionnaire constructed by the EJC.\textsuperscript{2} The questionnaire included sixteen questions covering the manner in which journalists use data, their levels of expertise, their learning needs, as well as the barriers they face (Bradshaw 2011b). The questionnaire was distributed through the School of Journalism and Mass Communication of Aristotle University to members of the Journalists’ Union of Macedonia and Thrace Daily Newspapers (ESIEMTH) and also to members of the Greek Chapter of the Open Knowledge Foundation. The survey was conducted from March until June 2014, and included 58 participants.

The questionnaire was given to journalists after a one and a half-hour lecture on data journalism, as part of a seminar held by the Educational Institution of Journalists...
Macedonia and Thrace Daily Newspapers. In the discussion that followed, two very interesting points emerged, which are captured in our research. One of the respondents to our questionnaire said that ‘there is not much data available in Greece in order to conduct analyses, to create visualizations into our articles, and that is very difficult to find those data’. Continuing the conversation, a journalist who took part in the survey added, ‘[e]ven if we find some data on the web, we do not know if we can use them because there is no license supplied with them’. Both comments agree with the results of the cities’ Open Data Census, conducted by the Open Knowledge Foundation Greece in 2014, in which it is shown that the main issues that come with data are the lack of licences and their non-machine-readable format, which makes them hard to be utilized by journalists.

**Survey results**

Of the 58 participants, 24 were men and 34 were women. The education level of the participants was quite high since 38 were graduates of journalism departments and sixteen had a master’s degree and four had a college degree in subjects relevant to journalism. Out of the 58 respondents, 32 were 25–35 years old, sixteen were 35–50 years old, eight respondents were in the age group of 18–25 years, and two respondents were over 50 years old. The largest percentage of participants worked in online media (55.2%; 32 participants). An impressive and rather expected fact because of the economic crisis in Greece is the 17.2% (ten participants) of journalists who are unemployed while the same percentage (17.2%) work with a newspaper, 3.45% (two participants) with a magazine and 6.9% (four participants) in the education sector.
When the participants were asked why they were involved in data journalism, 3.4% (two participants) cited competition as the reason, 20.7% (twelve participants) cited professional ambition, 6.9% (four participants) reported that they were curious, 27.6% (sixteen participants) reported that they wanted to develop a public interest service, and 41.4% (24 participants) said that they wanted to increase the productivity in news production. The results indicate that Greek journalists believe in the added value of data journalism both for the public good and for the production of news. Regarding the status of data journalism, the survey shows that 58.62% (34 participants) of our sample had no immediate plans to start working in data journalism, while 24.14% (fourteen participants) was planning to work with data in the next six months. An overall 10.34% (six participants) of the sample was currently working on a data project and only 6.90% (four participants) had already published a data journalism project.
Figure 1: Status of data journalism per publisher.

The stacked bar chart (see Figure 1) shows that, of the 58 participants, 3.45% (two participants), who work for a magazine, and 13.79% (eight participants), who work for a newspaper, have no immediate plans to start working in data journalism; 3.45% (two participants) of the participants, who are employed with a newspaper, are currently working on a data project; 27.59% (sixteen participants) of the participants, who work online, have no immediate plans to start working in data journalism; 17.24% (ten participants) are planning to work with data in the next six months; 3.45% are currently working on a data journalism project; and 6.90% have already published a data project. Of the unemployed among the 58 participants, 10.34% have no immediate plans to start working in data journalism, 3.45% are planning to work with data in the next six months, and 3.45% of them are currently working with data.
Among those who work in the education sector, 3.45% have no immediate plans to work with data, while 3.45% plan to work with data in the next six months. An interesting fact is that there is a percentage of unemployed who are currently working on data journalism projects, which may show that data journalism is considered a possible way out of unemployment.

Table 1 shows that all the participants currently employed with magazines have no immediate plans to start working in data journalism, while 20% (twelve participants) of the participants who work with newspapers are currently working on a data project and 80% (46 participants) have no immediate plans to start working in data journalism. Moreover, 12.5% (seven participants) of the participants who work online have already published a data project, 6.2% (four participants) are currently working on a data project, 31.2% (eighteen participants) are planning to work with data in the next six months, and 60% have no immediate plans to work with data. None of the unemployed journalists and those who are currently employed in education have published a data project. Among the unemployed journalists, 20% are currently working on a data project. The same proportion of unemployed are planning to work with data in the next six months and 60% (35 participants) have no immediate plans to work in data journalism.

<table>
<thead>
<tr>
<th>Cross-tab % within type of media</th>
<th>Have already published a data project (%)</th>
<th>Are currently working on a data project (%)</th>
<th>Are planning to work with data in the next six months (%)</th>
<th>Have no immediate plans to start working in data journalism (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magazine</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>
The results presented in Table 1 show that only journalists who work in online publications have already published a data project. It is worth noting that, except for journalists who are employed in magazines (and also account for a small percentage – 3.45 per cent of our sample), newspaper journalists seem to be aware of the data journalism phenomenon but they have no immediate plans to start data journalism projects. On the other hand, unemployed journalists seem to understand more the value of data journalism and appear to be more willing to start developing data journalism projects in comparison with newspaper journalists. Overall, the results agree with those of the corresponding research of EJC, which found that online journalists were the first to publish data journalism projects.

The barriers that journalists are facing for developing data journalism projects are grouped into six categories: lack of time, lack of resources, lack of adequate knowledge (e.g.: tools and working procedures), lack of adequate publishing infrastructure (e.g.: CMS fit for standard posts only), lack of management support, and lack of interest from staff. The percentage of participants citing lack of time was 48.28% (28 participants), those citing lack of resources constituted 72.41% (42 participants), those reporting lack of adequate knowledge constituted 72.41%,
participants reporting lack of adequate publishing infrastructure constituted 34.48% (twenty participants), those reporting lack of management support represented 3.45% (two participants), and participants citing lack of interest from staff represented 27.59% (sixteen participants).

Particularly noteworthy are the high levels in both lack of resources and lack of adequate knowledge. Lack of resources may be explained by the economic crisis in Greece that prohibits spending money on new resources. Although data journalism cannot be considered highly expensive, it certainly needs money and time to be spent on journalists’ training.

Figure 2: Barriers for data journalism per data journalism status (involvement in data journalism projects).

Figure 2 demonstrates the barriers for data journalism per data journalism status (involvement in data journalism projects). As expected, the barrier percentages are higher in the category with no immediate plans and lower in the category that has
already published a data project. Specifically in the category with no immediate plans, lack of time barrier represented 27.59% of participants (sixteen participants), lack of resources percentage was a high 48.28%, lack of adequate knowledge was reported by 37.93% (22 participants), lack of adequate publishing infrastructure represented 24.14% (fourteen participants), lack of management was reported by 3.45%, and lack of interest from staff constituted 10.34% (six participants). On the other hand, in the category that had already published a data project the barrier percentages ranged from 0% to 8%. These results can be justified by the fact that journalists who have already been involved in data journalism projects have confronted these barriers and they have found ways to overcome them.

Next the survey examined the journalists’ data needs. Specifically, journalists were asked to evaluate the importance of raw data as a primary source in journalism, the skills to analyse data in journalism, and their interest in acquiring the skills to practise data journalism. The results of this evaluation are included in Table 2. Raw data were considered important or very important by 89.7% (52 participants) of the journalists. An overall 72.4% (42 participants) of journalists believed the same for the skills for analysing data. Regarding the skills for practising data journalism, 86.2% (50 participants) considered them rather important or very important. Overall, the majority of journalists acknowledged the importance of raw data and skills in practising data journalism.

<table>
<thead>
<tr>
<th></th>
<th>Raw data (%)</th>
<th>Skills to analyse data (%)</th>
<th>Skills to practise data (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Of little importance</td>
<td>3.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.9</td>
<td>27.6</td>
<td>13.8</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Moderately important</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Important</td>
<td>34.5</td>
<td>37.9</td>
<td>41.4</td>
</tr>
<tr>
<td>Very important</td>
<td>55.2</td>
<td>34.5</td>
<td>44.8</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of journalists’ data needs.

It is encouraging, as seen in the survey, that 58.62% (34 participants) of the journalists frequently or very frequently used public data and statistics from local, European and/or international agencies and organizations (such as Eurostat, etc.) as sources for reporting. Nevertheless, there is still a large amount of data needed, on government issues (reported by 21% of the participants [twelve participants]), on the economy (17% [ten participants]), for research (15% [nine participants]), on health (13% [eight participants]), on education (12% [seven participants]), on crime (10% [six participants]), for science (10% [six participants]), and on budget (2% [one participant]).

When journalists were asked to envision their organization in order to start engaging in data journalism, 62.1% (36 participants) proposed a combination of external experts and existing staff. Also 31% of the participants suggested training existing staff, and 6.9% (four participants) answered that data journalism was not an option for their organization.

Further, journalists answered that when their journalistic work required programming they collaborated with programmers (62.1%, 36 participants), or they took care of the programming on their own (13.8%, eight participants). A percentage of journalists
said they never needed programming in their work (24.1%, fourteen participants). With regard to the analysis of data for reporting, most of them reported that they did the statistical analysis on their own (41.4%, 24 participants), 31% (eighteen participants) collaborated with statisticians, and 27.6% (sixteen participants) never needed statistical data analysis in their work. Finally for the production of data visualization for their stories they either collaborated with a visual designer (27.6%), or produced the data visualization on their own (34.5%, twenty participants), or they never published data visualizations with their stories (37.9%, 22 participants).

<table>
<thead>
<tr>
<th>Programming needs</th>
<th>Data analysis needs</th>
<th>Data visualization needs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collaborate with others</td>
<td>62.1</td>
<td>31.0</td>
</tr>
<tr>
<td>Do it on your own</td>
<td>13.8</td>
<td>41.4</td>
</tr>
<tr>
<td>Never needed</td>
<td>24.1</td>
<td>27.6</td>
</tr>
</tbody>
</table>

**Table 3**: How journalists cope with their programming, analysis and visualization needs.

The majority of journalists who participated in the survey (79.3%, 46 participants) were interested in being trained in integrating data into stories (e.g., giving more depth to a story and providing context through the use of statistical data and data visualization), 41% (24 participants) were interested in using computing techniques to analyse data sets on a deeper level, and 62% (36 participants) responded data as
service (e.g., developing applications for data reporting or other services for readers; *New York Times* ‘Is it better to buy or rent?’).

Figure 3: Journalist interests in learning per medium.

Finally, Figure 3 includes the survey participants’ interests in learning per medium. Only 3.45% (two participants) of journalists working in magazines were interested in learning data visualization, while they exhibited no interest in any other subject relating to data journalism. Newspaper journalists seemed to be slightly more interested in learning, with 13.79% (eight participants) being interested in learning ways to find data. The next group with higher learning interests was the unemployed, who appeared to be more motivated with the visualization of data (17%, ten participants). Online journalists were the most interested in learning issues related to data journalism, with all the topics ranging from 24% to 38%. Finally it is worth mentioning that journalists working in education were interested equally in finding data, checking whether the data were reliable, cleaning, analysing and visualizing data, developing stories from data and integrating data projects with publishing systems.
Discussion of the findings

The results of the survey indicate that the majority of Greek journalists are not involved in data journalism projects. Journalists working in online media seem to have an edge on data journalism in comparison with journalists employed in other media or in the education sector. Nevertheless, the majority of journalists agreed with the importance of working with data and were interested in acquiring more knowledge and skills concerning data journalism. It is worth noting that online journalists appear to be more involved with data journalism and keen to improve their knowledge and skills in data journalism. This result agrees with the result of the EJC survey. But we must take into account the fact that the percentage of journalists working in magazines is quite small and thus the results concerning their attitude towards data journalism may not be totally accurate.

Journalists believe that they need more data on various issues. The requirement for data on government issues, economy, education and health needs is high for in-depth research. The journalists remarked that financial transparency can make the Greek government and citizens accountable for their actions and help fight corruption in Greek society. It is quite obvious that the amount of data available on the web is growing rapidly. Thus, if journalists want to stay competitive, they must adapt to the demands of their profession and become, to some extent, data journalists.

It is obvious that newspaper journalists in contrast with online journalists have failed to adapt to the data journalism era. The migration from newsprint to online news is a worldwide phenomenon (Cokley et al. 2011; Cokley et al. 2015). The significant involvement of online journalists in data journalism and their increased interest to
acquire more skills in order to produce more and better data journalism projects is evident from the survey. Overall, Greek journalists need training on data journalism issues. In the next sections we will attempt to clearly define the necessary skills that journalists must have in order to cope with data journalism.

**Defining data journalism**

At the end of the twentieth century, employing a large amount of data to write an article was difficult and required skills that went beyond the capabilities of the average journalist. Some news organization in the United States and Great Britain even hired programmers that worked on novel news products (Parasie and Dagiral 2013). Journalists used to rely on information provided by various sources (governments, officials, research studies, etc.). Of course there were some cases of investigative journalism where journalists were able to find resources to gather and analyse their own data and publish their results in articles. But the situation changed rapidly as a growing amount of data gradually became available online and efficient tools with which anyone can analyse, visualize and publish large amounts of data appeared (Sirkkunen 2011).

The concept of data journalism in not new. Digital data have been utilized in news production since the late 60s in US newspapers (Parasie and Dagiral 2013). Data journalism gradually emerged with the rapid introduction of ICTs and the availability of data in digital form. It is synonymous with the term data-driven journalism while the older term, computer-assisted reporting, has vanished (it was introduced in the early stages of computer history) (Bradshaw 2010). In data journalism there is an
increased interaction between journalists and several other fields such as design, computer science and statistics (Thibodeaux 2011).

The term data journalism is attributed to Simon Rogers, who first mentioned it in a post to the Guardian Insider Blog (Knight 2015). **Data journalism is a field that combines spreadsheets, graphics data analysis and the biggest news stories (Rogers 2008).** It can be viewed as a **process** that begins with analysis, and continues with filtering and visualizing data in a form that links to a narrative (Lorenz 2010). It is fundamentally the production of news graphics and includes elements of design and interactivity (Bradshaw 2010; Lorenz 2010; Rogers 2008). Megan Knight (2015) describes data journalism as ‘a story whose primary source or “peg” is numeric (rather than anecdotal), or a story which contains a substantial element of data or visualization’. The concepts of traditional journalism and data journalism are depicted in Figure 4.
Figure 4: Traditional versus data journalism concepts.

Based on the above and in order to better address the power of visualization and interactivity that are significant factors in data journalism we propose that data journalism be defined as the process of extracting useful information from data, writing articles based on the information, and embedding visualizations (interacting in some cases) in the articles that help readers understand the significance of the story or allow them to pinpoint data that relate to them.

**Journalists’ ICT skills**

Today the journalist is expected to possess the necessary ICT skills to cope successfully with the challenges in his everyday work. Journalists today often seek
information on the web and by e-mail (Veglis 2013). Writing news articles, constructing diagrams via spreadsheet applications, communication via e-mail, visualizing data with the help of various applications, publishing material on the WWW are some common examples of skills that journalists need to possess (Peebles 2011). It is worth noting that these skills must be sustained and extended by adopting news tools, services and applications as they become available. Veglis and Pomportsis (2012, 2014) organized the journalists’ ICT skills into five categories: basic skills, web publishing skills, Web 2.0 skills, web casting skills and data journalism skills. This categorization has to be extended by adding an additional category that refers to Web 3.0.

Web 3.0 skills include basic knowledge and experience with Web 3.0 technologies. Two of the most important components of Web 3.0 for the journalist are the Semantic Web and Linked Open Data. The Semantic Web is an initiative that aims at promoting common data formats on the WWW by encouraging the inclusion of semantic content in web pages. By this way the web, which is currently dominated by unstructured and semi-structured documents, is converted into a web of data. It is based on the Resource Description Framework (RDF) that provides a common framework for data to be shared and reused across applications, enterprises and communities (Allemang and Hendler 2011). Moreover; a data journalist should be able to retrieve the semantic data using SPARQL Protocol and RDF Query Language (SPARQL) (Prud’hommeaux and Seaborne 2013). Depending on the type of semantic data that a journalist may require to investigate, he may need to be familiar with specific vocabularies and ontologies that are described using Resource Description Framework Schema (RDFS) (Brickley and Guha 2014) or Ontology Web Language (OWL) (McGuinness and van Harmelen 2014). Such vocabularies are utilized in the
web of data or Linked Open Data (Bizer et al. 2009) and are described in the Linked Open Data Vocabularies of the Open Knowledge Foundation as a living ecosystem (Vatant and Vandenbussche 2014). Open data is a movement that promotes the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control (Auer et al. 2007). According to the Open Data definition, Open Data are data that can be freely used, reused and redistributed by anyone, subject only, at most, to the requirement to attribute and/or share-alike (Dietrich et al. 2012). Open Data may be accompanied by an open licence like ODBL \(^4\) (more appropriate), or Creative Commons,\(^5\) which assigns copyright or licenses third parties to freely reuse the data. The linked open data are a crucial parameter of the Web 3.0 since they can be linked to other data and thus support the semantic web.

**Summarizing, the journalists must be able to:** understand the basics of Web 3.0; use the limited Web 3.0 tools currently available (but expected to grow exponentially in the near future); and create articles that comply with open data rules and link to other open data.

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, developed a 5-star deployment scheme for Open Data in order to encourage people – especially government data owners – to produce good linked Open Data (2010). This system can be adapted for journalists as a guide for them to produce articles in the Web 3.0 era. The proposed star rating system is presented in table 4.

<table>
<thead>
<tr>
<th>Open Data star</th>
<th>Benefits for data journalism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td></td>
</tr>
<tr>
<td>Data stars</td>
<td>classification</td>
</tr>
<tr>
<td>------------</td>
<td>----------------</td>
</tr>
<tr>
<td>★</td>
<td>data are accessible on the web under an open licence (such as PDDL,ODC-by or CC0); however, the data are locked up in a document.</td>
</tr>
<tr>
<td>★★</td>
<td>available as structured data (e.g., Excel)</td>
</tr>
<tr>
<td>★★★</td>
<td>non-proprietary formats (e.g., CSV instead of Excel)</td>
</tr>
<tr>
<td>★★★★</td>
<td>data items have a Uniform Resource Identifier URI(^1) and can be shared on the Web.</td>
</tr>
</tbody>
</table>
The data are not linked to other related data. However, the data are linked to other data to provide context (Linked Open Data) to help the data journalist discover more (related) data and combine it in his research, reuse it in the statistical analysis and in the visualizations of his story.

Table 4: Star rating system for the use of good linked open data for journalists.

[1] It is a string of characters used to identify a name of a resource over the WWW. The most common form of URI is the uniform resource locator (URL), frequently referred as a web address.

Data journalism stages – skills

There are two fundamental approaches to creating a data journalism article. In the first approach, a data set provides more information for a story topic that was already discovered by the newsroom, whereas in the second a data set serves as a starting point for finding additional information (Sirkkunen 2011).

Mirko Lorenz, an information architect and multimedia journalist, proposes the stages of data journalism workflow: digging deep into data by scraping, cleansing and structuring it, filtering by mining for specific information, and visualizing and making...
a story (Lorenz 2010). Paul Bradshaw describes the process of data journalism in a similar manner: data must be found, which may require specialized skills like MySQL or Python, then interrogated, for which understanding of jargon and statistics is necessary, and finally visualized and mashed with the aid of open source tools (2010). So data journalism can be viewed as a process of refinement, where raw data are transformed into something meaningful and as a result the value to the public grows, especially when complex facts are boiled down into a clear story that people can easily understand and remember (European Journalism Centre 2010). Also Paul Bradshaw (2011a) proposed the inverted pyramid of data journalism. The inverted pyramid includes five stages: Compile, Clean, Context, Combine and Communicate.

Based on what has been reported we propose a data journalism workflow that includes six stages: Data Compilation, Data Cleaning, Data Understanding, Data Validation, Data Visualization and Article Writing. The workflow model is depicted in figure 5. These stages can indicate the skills that journalists must have to be data journalists. Next we discuss each of the six stages of the workflow.

![Data Journalism Workflow](image)

**Figure 5:** Data journalism workflow.

*Data compilation:* data journalism begins in one of two ways: either the journalist has a question that needs data, or a data set that needs questioning. Whichever it is, the compilation of data can take one of the five following forms: (1) data may be supplied directly by an organization (in some cases in the form of open data); (2) data may be
found with the help of advanced searching techniques; (3) data may be compiled by scraping web pages; (4) data may be collected by converting documents to other formats that can be analysed; and (5) data may be collected by means of observation, surveys, online forms or crowd-sourcing (Bradshaw 2011a). It is worth noting that data scraping is a process in which a software tool extracts data from human-readable output coming from other software – for example, using the command ImportHTML command in Google spreadsheet in order to scrape a table or a list from a web page. Also there are free tools on the Internet that allow users to convert documents to other forms that can be analysed. For example, DocumentCloud is a web-based software platform created specifically for journalists to allow the searching, analysing, annotation and publication of primary source documents used in reporting.

Data Cleaning: also known as data scrubbing, this is the process of detecting and correcting (or removing) corrupted or inaccurate records from a data set (Wu 2013). The process usually takes two forms: removing human errors and converting the data into a format that is consistent with other data the journalist is using. Typical examples include empty entries, duplicate entries, the use of default values to save time or where no information was held, incorrect formatting (e.g., words instead of numbers), corrupted entries or entries with HTML code, multiple names for the same thing and missing data (Bradshaw 2011a). Cleaning data can be done in simple ways: for example, using find and replace commands or filters on spreadsheets. Of course there are specialized tools like Google’s OpenRefine, which is a stand-alone open-source desktop application for data cleanup and transformation to other formats. The interface resembles spreadsheet applications and it can handle spreadsheet file formats.
**Data Understanding:** data sets usually include various codes that represent categories, classifications or locations, and special terminology that is not easy to be understood by journalists. Also in many cases further data are needed for existing data to become meaningful. That means that journalists ought to have the ability to consume knowledge, produce coherently and think critically about data. This ability is referred to as data literacy. Data literacy also includes statistical literacy but also understanding how to work with large data sets, how they were produced, how to connect various data sets and how to interpret them (Gray et al. 2012).

**Data Validation:** in this stage, the journalist cross-checks his original data and obtains further data from sources in order to enrich the available information (Silverman 2014; Veglis 2013). Like any source, data cannot always be trusted since it comes with its own histories, biases and objectives. Thus the journalist has to investigate issues like: who gathered it, when, for what purpose, and how was it gathered (Bradshaw 2011a). This can be accomplished by investigating the history of the creation of the data set, by finding references to the data set or by using other sources of information that refer to the same subject he investigates (Silverman 2014).

**Data Visualization:** data visualization is a modern branch of descriptive statistics that involves the creation and study of the visual representation of data. It is the graphical display of abstract information for data analysis and communication (Cairo 2012). The visualization can be static, thus offering only pre-composed ‘views’ of data, best suited for static medium (e.g. print), or it can be interactive, thus supporting multiple static views in order to present a variety of perspectives on the same information. In interactive data visualization there is a user input (a control of some aspect of the visual representation of information) and the changes made by the user must be incorporated into the visualization in a timely manner (Veglis 2018). It is worth
mentioning that infographics are also part of the static visualization. They are graphic
visual representations of data or knowledge which are able to present complex
information quickly and clearly (Smiciklas 2012). They are often used in newspapers,
to show the weather, as well as in maps, site plans, and graphs for statistical data.

Article Writing: the last stage includes the writing of the news article. Depending on
the intended publication medium, the article may include special characteristics (e.g.,
external links to other articles or related material, multimedia content, mashups, and
static or interactive visualizations) to fully exploit the medium’s potential.

From the previous discussion it is clear that in order for the journalist to be able to
cope successfully with all stages of data journalism he must have very specialized
ICT skills. That is the reason why the majority of news organizations that utilize data
journalism have created teams of different specialities in order to publish articles
based on data sets or launch data-driven news apps (Gray et al. 2012). But even if the
journalist is working in a team that includes programmers and data visualization
experts, the previously described skills will enable him to collaborate effectively with
the members of his team, thus resulting in successful data journalism projects.

Conclusions and future work

This article discusses the issue of data journalism. Special consideration has been
given to the necessary skills that journalists should have in order to cope with this
new journalism trend. These skills are closely related with the data journalism stages
that have been described in great detail.

The above findings can guide journalism educators in adapting their programs. This is
already happening since the constant changes in market demands, as far as journalistic
skills are concerned, have forced journalism educators to adjust their programs to better facilitate the needs of the industry (Wenger and Owens 2012). This way, journalism graduates will enter the media market better prepared to cope with their work demands (Veglis 2013).

A survey concerning the use of data journalism in Greece was also conducted. Its results are in agreement with a similar survey conducted in Great Britain, which also described a low penetration of data journalism practices in media organizations (Knight 2015). There is no doubt that data journalism requires special skills on the part of journalists. The luck of such skills is evident in the case of Greece. The majority of journalists are not involved yet in data journalism projects, although they are willing to work on such projects in the future. Online journalists appear to be more involved and more motivated and this finding is in agreement with studies that indicate that there is an ongoing shift to online journalism (Cokley et al. 2015). We must not forget that the full potential of data journalism can only be exploited on the web.

The results of this study can help journalist educators create training programs on specific skills that relate to data journalism. It is worth noting that the School of Journalism & MC of Aristotle University of Greece will start offering life-long learning programs on data journalism specifically targeting professional journalists (Veglis and Bratsas 2015).

Tim Berners-Lee believes that ‘data-driven journalism is the future’ and urges reporters to hunt for stories in data sets (Arthur 2010). We are convinced that the Semantic Web and Linked Open Data will play a significant role in the evolution of data journalism. In the future journalists should be able to comprehend and utilize advanced technologies that will include more ‘intelligent web’ (Bradshaw and
Rohumaa 2011). Semantic Web technologies, smart devices (smartphones, etc.) and tools are continuously being transformed and upgraded. Data journalism ought to adopt these technologies and thus data journalists need to acquire the necessary skills in order to exploit them. Future extension of this work will include a thorough large-scale survey on the journalists’ educational needs as far as data journalism skills are concerned.

References


Accessed 20 April 2015.


**Contributor details**

Andreas Veglis is professor and head of the Media Informatics Lab at the School of Journalism & Mass Communication at the Aristotle University of Thessaloniki. He received his BS.c. in Physics, MS.c. in Electronics and Communications, and Ph.D. in Computer Science, all from Aristotle University. His research interests include information technology in journalism, new media, course support environments, data journalism, user-generated content, verification, fake news and open data. He is the author or co-author of twelve books. He has published 63 papers on scientific journals and has presented 97 papers in international and national conferences. Professor Veglis has been involved in 25 national and international research projects.
Dr Charalampos Bratsas is the founder of the Greek Chapter of the Open Knowledge Foundation and the President of its Administration Board. He has extensive experience in Semantic Web with a focus on Linked Open Data Internationalization and Ontology engineering. He has been involved in ten EU research projects. He is currently a faculty member of the master in web science with the Mathematical Department of the Aristotle University of Thessaloniki. His main areas of interest are: Data Mining and algorithms, Semantic Web, Social Semantic Web, Internationalization of Linked Data, Applications of Artificial Intelligence, Neuroinformatics, as well as, Linked Education. He is an author in two theses, six book chapters, eleven articles in journals, and more than 30 articles published in International conference proceedings.

Contact:

Aristotle University of Thessaloniki, 54126, Thessaloniki, Greece.

E-mail: veglis@jour.auth.gr

E-mail: cbratsas@math.auth.gr

Notes

1 http://datajournalism.okfn.gr/

2 http://ejc.net/.
3 More information on the results of the survey are available at
http://www.slideshare.net/lilianabounegru/preliminary-results-of-ejc-survey-on-
training-needs-for-data-journalism.

4 http://opendatacommons.org/licenses/odbl.

5 http://creativecommons.org/.


7 https://support.google.com/docs/answer/3093339 ?hl=en#3093339.

8 http://www.documentcloud.org/.

9 http://openrefine.org/.