social sciences

MDPI

Review

Applications of Big Data in Media Organizations

Andreas Veglis *, Theodora Saridou, Kosmas Panagiotidis, Christina Karypidou and Efthimis Kotenidis

Media Informatics Lab, School of Journalism & Mass Communication, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

* Correspondence: veglis@jour.auth.gr

Abstract: The exploitation of data in the media industry has always played a significant role. This is especially evident today, since data (and in many cases big data) are generated through various activities that relate to the production and also consumption of news. This paper attempts to highlight the importance of big data utilization in the media industry. Specifically, it discusses cases of big data exploitation, such as media content consumption and management, data journalism production, social content utilization, and participatory journalism applications. The study also examines the changes that big data has introduced in all stages of the journalism practice, from news production to news distribution, by utilizing the available tools. Finally, it discusses new developments that relate to semantic web (Web 3.0) technologies, which have already started to be adopted by media organizations around the world.

Keywords: media organizations; social media; user-generated content; participatory journalism; data mining; semantic web technologies

Citation: Veglis, Andreas, Theodora Saridou, Kosmas Panagiotidis, Christina Karypidou, and Efthimis Kotenidis. 2022. Applications of Big Data in Media Organizations. *Social Sciences* 11: 414. https://doi.org/10.3390/socsci11090414

Academic Editor: Javier Díaz-Noci

Received: 14 August 2022
Accepted: 6 September 2022
Published: 8 September 2022

1. Introduction

In the last 20 years, the amount of digital data that is being produced by various human activities is increasing at an exponential rate (Veglis and Maniou 2018). The need for handling big data is considered to be a dominant factor due to the added value one can gain through working with data. We are living in the era of big data, which is huge amounts of data in digital form. The study of big data can offer interesting insights into numerous scientific disciplines and industries (Kitchin 2014; Kitchin and McArdle 2016).

The exploitation of data in the media industry has always played an important role, especially nowadays, when people interact with various sources of information and spend more time online, producing data through their devices (smartphones, tablets, laptops, etc). The process of analyzing and interpreting these data, in terms of profit, is becoming more productive than ever before for the media sector (Nelson and Webster 2016; Stone 2014). Data journalism is another area that directly relates to the exploitation of big data. This journalism specialty has gathered significant attention in the last decade, and it is considered to be a significant factor in the development of the media industry (Veglis and Bratsas 2017). Social media, and specifically, User-Generated Content (UGC), is a valuable source of information for media organizations (Veglis 2014). Every day, significant amounts of UGC are generated. UGC is also related to participatory journalism (Saridou and Veglis 2016), which is believed to be an area that needs to be developed more by media organizations. The sheer volume of UGC means that it can be classified as big data. Based on the above, it is quite obvious that many areas that relate to the media sector, have to do with big data.

This paper aims to highlight the importance of big data utilization by media organizations. It examines all the previously mentioned areas of big data exploitation, namely, media content consumption and management, data journalism investigation and production, the role of data mining in big data applications, social content utilization and participatory journalism applications. It thus focuses on case studies from the media field that employ

big data strategies in order to monitor user-generated content posted both on the company website and on social media platforms. The paper also explores whether and how big data can be used to minimize the risks and the problems of UGC, as it happens with customers by brands in marketing. It also discusses the developments that big data has introduced in all the stages of the journalism practice, from news production to news distribution, by utilizing the available tools. Finally, the study examines the changes that semantic web technologies, in relation to the exploitation of big data, have already started to implement in the media sector.

Based on the above we can conclude that the main objectives of the study are:

- Identify the importance of big data utilization by media organizations.
- Present the application areas of big data exploitation (media content consumption and management, data journalism investigation and production).
- Highlight the role of data mining in big data applications, social content utilization, and participatory journalism applications.
- Discuss the changes that semantic web technologies will introduce to the exploitation of big data in media organizations.

2. Big Data in Media Organizations

The term "big data" was proposed at the end of the 20th century. It portrays datasets so large that they cannot be captured, curated, managed, and processed by commonly used software running on standard personal computers (Lewis and Westlund 2015; Snijders et al. 2012). Kitchin (2014) proposed a more detailed definition: "Big Data is huge in volume (terabytes or petabytes), high in velocity (being created in or near real-time), diverse in variety (structured and unstructured in nature), exhaustive in scope (striving to capture entire populations or systems), fine-grained in resolution and uniquely indexical in identification, relational in nature (containing common fields that enable the conjoining of different data sets), and flexible (can be extended and expanded)."

Big data has had an impact on many industries, including the media industry (Veglis and Maniou 2018), where its application has been facilitated by new technological developments that have automated and, to some extent, simplified data analysis (Stone 2014). Big data exhibit the following characteristics: volume, variety, velocity, and veracity (Hilbert 2016).

Media organizations realized that by studying content consumption data, they can extract useful information which may help in designing successful publishing strategies and lead to new revenue opportunities. Moreover, with the digitization of the journalistic process and the introduction of the internet and its services, media organizations have at their disposal multiple publishing channels (Veglis 2012). Each one of these channels represents a source of data, and all of them together constitute large volumes of information that can be categorized as big data. Newspaper circulation, Radio and TV ratings, audience views, searches, preferences, clickstreams, log files, and social media sentiment are some examples of data provided by the channels mentioned above and which can be exploited by media organizations (Newman et al. 2018; Stone 2014).

Big data can also be used in the production process in media organizations. Specifically, the availability of data in digital form and the abundance of efficient online tools that analyze, visualize, and publish large amounts of data have fueled the introduction of data journalism (Veglis and Bratsas 2017). Data journalism can be defined as the process of extracting useful information from data, writing articles based on the information, and embedding visualizations (interactive in some cases) in the articles that help users understand the significance of the story or allow them to pinpoint data that relate to them. Thus, big data encourage the use of infographics and data visualizations in journalistic projects and introduce new facets for understanding the transformation of raw information into journalistic truth (Kalatzi et al. 2018; Karypidou et al. 2019).

Social media content, especially user-generated content, is considered to be a valuable source of information for journalism organizations (Veglis 2014). However, collecting, archiving, and exploiting such data is not an easy task, since the amount of data and their

diversity require special skills and increased computational resources. Lastly, big data is also employed in the case of participatory journalism, where journalists are called to exploit the information provided by the public (Saridou and Veglis 2016).

Taking into consideration the multiple uses of big data in media organizations and in order to have a better understanding of their dynamics in different application areas, all the information was included in a visualization. As shown in Figure 1, media organizations can take advantage of big data for content consumption and management, data journalism applications and practices, and data mining purposes.

Figure 1. Application areas of big data in media organizations.

Next, the previously mentioned areas, where big data have a significant impact on the operation of media organizations, are being analyzed.

3. Content Consumption

Since the digitalization of the media industry and the evolution of ICT, there has been a fundamental change in the media industry on how it uses data and analytics. Media organizations utilize big data to understand why users subscribe and unsubscribe to their services (Evens and Van Damme 2016). During the last decades, the way news is consumed by people has fundamentally changed. The convergence of information and communication technologies (Spyridou et al. 2013), along with the proliferation of the internet services (Veglis and Pomportsis 2014) and the availability of a wide range of portable devices (Thurman et al. 2018) has turned readers, listeners, and viewers into users who hold a decisive role in content consumption (Turnbull 2020). According to the latest Reuters Institute Digital News Report, the use of smartphones for news has grown at its fastest rate for many years, especially during Coronavirus lockdowns, while the use of laptops, desktop computers, and tablets is stable or falling (Newman et al. 2021). In this vein, media organizations adapt their news production process by using, for example, native taps and swipes to break up narratives and by providing visually rich formats that reshape storytelling in the mobile context (Newman et al. 2018).

Apart from different devices, news can also be found in a plethora of online sources, breaking the exclusivity of mainstream media outlets. The web and alternative news sources have offered several opportunities for varied content consumption, which is additionally influenced by the collapse of local, daily newspapers and the growing 24-h cable news cycle (Bentley et al. 2019). People can now discover stories from thousands of websites, social

media platforms, email newsletters, or direct browsing (Bentley et al. 2019). Specifically, research suggests that social media are becoming central to the way people experience news, as networked media technologies are extending the users' ability to create and receive personalized news streams (Hermida et al. 2012). A large majority of users visit one or more social networks or messaging apps for consuming, sharing, or discussing news. The largest proportion of social media news users says they are most likely to pay attention to mainstream media and journalists, although there are many who appreciate the alternative perspectives as well (Newman et al. 2021).

Consumer data collected from social media user behavior often reveal overlooked factors that have the potential to drive consumer interest. Big data help media organizations to understand the demand for different types of news and types of content for a given age group on different available publishing channels. They also investigate when users are most likely to view content and on what devices. The data sizes that the major players in the media sector are employing today are quite staggering. In 2017, Facebook collected and processed 500 Terabytes of data every day (Singh 2017).

Within this framework, consumption becomes fragmented, since many media outlets face the challenge of customizing their content or programming in order to appeal to different niche audiences (Gruszynski Sanseverino and De Lima Santos 2021). The significant growth in the use of social networking platforms and mobile devices for news access has facilitated the exposure to personalized information spaces (Thurman et al. 2018). As a result, news organizations that pursue reader loyalty and trust have a strong incentive to implement personalization algorithms, aiming to achieve particular goals by taking into account diverse user attitudes and providing high-quality recommendations (Bodó et al. 2019). For media organizations, big data strategies can include audience analytics to enable a deeper understanding and targeting of users, tools to utilize public and private databases for journalistic storytelling, tools to manage and search the exploding amount of content, and tools to automate the production of text or video stories (Stone 2014).

In the light of such widespread changes, research evidence underscores data exploitation practices by several organizations. For example, Huffington Post uses big data in order to optimize content, ensure the efficacy of native advertising, regulate advertising placement, and create passive personalization. The mainstream media outlet pursues a more accurate analytical approach to decision making in order to improve user and advertiser experience, while enabling content delivery at the right time, on the right device, and to the right audience (Stone 2014). Similarly, the New York Times aims to customize online news delivery by adjusting users' experience to accommodate individual interests through an enriched experience that keeps the most important and compelling news at the center of the website for everyone but treats readers according to their unique preferences and habits (Spayd 2017). In the same context, the Nielsen Company recently introduced its new measurement service, which utilizes vast sets of data mined from Twitter. Although data have always played an important role in the television industry, particularly in relation to ratings and market research, it is the volume, variety, and velocity of big data that have forced the development of new innovative practices (Kelly 2019).

4. Data Journalism

In the last 30 years, digital technologies with the introduction of various tools have made journalistic work easier. However, they have also made journalist work more difficult, because they have overwhelmed journalists with more information than can be handled by their investigative toolboxes (Venturini et al. 2018). Data journalism emerges as a result of these changes, and it is related to data-driven journalism. Specifically, the introduction of Information and Communications Technology (ICT) and the availability big data have turned data journalism into its current form. Big data are sociocultural phenomena that differentiate data journalism from other forms of journalism that have used huge amounts of data or databases (Sandoval-Martín and La-Rosa 2018). Bradshaw (2010) marks that

"data can be the source of data journalism, or it can be the tool with which the story is told, or it can be both".

As a term, data Journalism was first referred to by Rogers Simon in a post to the Guardian Insider Blog (Knight 2015). According to him (Rogers 2014), data journalism promotes open journalism and open data. The term open data is related to transparency; accountability; accessibility; and free, public, and recyclable use.

This kind of journalism should be approached not as a technology that needs to be adopted or as a given existing practice, but rather, as something that "materially and incoherently exists in a fundamentally relational space, across organizations, outside the news organizations, and even possibly across the national framework", as De Maeyer et al. (2015) noticed. It is important to clarify that data journalism is a type of journalism that follows the same fundamentals as all the other types (investigative journalism, reporting, etc.) in order to build a good story, but the main difference is that the story is based on information that comes from data and not from the traditional sources (Kalatzi et al. 2018). This way of writing is characterized by complex work and collaboration methods between specialists and academics (Charbonneaux and Gkouskou-Giannakou 2015).

Visualizations are an integral part of data journalism articles and are employed in order to convey large amounts of data as meaningful information (Veglis and Bratsas 2017). Data Journalism articles are enriched with statistical data and comprehensive visualizations, which allow news reporters to publish stories with complex data acquirable from wherever they are (Hahn and Stalph 2016). Data visualization is a great job today (Cubbit 2015).

Visualizations can be static or interactive. In a static visualization, there is only one view of data, and on many occasions, multiple cases are needed in order to fully understand the available information (Veglis 2009). Interactive visualizations can empower people to explore data on their own. An interactive visualization should initially offer an overview of the data, but it must also include tools for discovering details. It may also include animated transitions and well-crafted interfaces in order to engage the audience with the subject it covers (Murray 2017). Thus, data visualizations can enhance storytelling and help to convey complex topics (Hahn and Stalph 2016).

In journalism, big data should be open. Big data and related approaches present new perspectives for understanding the epistemology of converting the raw information into journalistic truth. With regard to news distribution, big data is associated with emerging presentations of digital journalism, such as infographics, interactive data depictions, and adaptable probability models, among others (Lewis and Westlund 2015).

Thus, data journalism mainly relies on big data. Panama papers and WikiLeaks are some widely known examples in which the availability of big data leads to significant journalistic successes. However, in order for the media organization to be able to utilize big data in terms of cleaning, understanding, validating, and visualizing, a significant variety of skills are required (Veglis and Bratsas 2017). In the big data era, a new working model is being observed to be developing in media organizations: collaborative work. Thus, journalists, programmers, web developers, and designers work together as a team in the news production process (Sandoval-Martín and La-Rosa 2018).

Traditional journalistic methods are mixed with data analysis, programming and visualization techniques (Appelgren and Nygren 2014). Thus, the working groups consist of a combination of skills in journalism, web development, data analysis, visualization and statistics. In this context, new departments in media organizations, such as The New York Times or The Guardian, as well as independent organizations, such as ProPublica.org, and less formal groups of investigative journalists who have published articles based on data processing techniques have been observed (Parasie 2015).

Currently, the majority of media organizations around the globe are developing data journalism projects. A recent study (Karypidou et al. 2019) surveyed six major media organizations (The Guardian, The New York Times, the BBC, the CNN, the Associated Press, and the Reuters) and found the significant effort that is being invested toward the development of data journalism. Nevertheless, although everybody understands the

importance of data journalism, the latter is not yet accepted as mainstream journalism, and rarely data journalism articles are published as main stories.

5. Audience Participation

Although audience participation has always been part of the journalism practice, the diffusion of Web 2.0 tools along with the socio-economic circumstances have led to the proliferation of user-generated content (UGC) and increased users' involvement in the news production process. In order to respond to the changing conditions, media organizations have gradually redefined their work strategies, adopting participatory formats that allow readers to actively consume or co-produce content (Tong 2015). The adoption of amateur content is thus combined with editorial control, under a professional umbrella (Deuze 2006). Users contribute to the ordinary news practice through content rating, polls, sharing to social networks, submission of audiovisual or textual material, collaborative content, comments, discussion forums, and citizen blogs (Spyridou 2018). Audience participation in news production can be enabled by data journalism projects as well. Thanks to the participation of victims and witnesses, a number of media organizations in Latin America have revealed situations involving huge breaches of human rights not identified in official records (Palomo et al. 2019). In Italy, data journalists specifically rely on the contributions of users, which are seen as co-constructors of reality (Young et al. 2018). Apart from news websites, a wealth of data is also produced on social media platforms, where journalists look for breaking news events, find ideas for stories, keep in touch with their audience, and collect information (Weaver and Willnat 2016).

However, the integration of participatory journalism practices into professional routine can raise a wide range of problems, aptly characterized as dark participation (Quandt 2018). Journalists often underline the excessive use of inappropriate language, flaming, stereotyping, and superficial discourse by the participants (Manosevitch 2011), while incivility is recorded as a common obstacle in comment sections (Ksiazek et al. 2015). Furthermore, the examples of dark participation range from misinformation and hate campaigns to individual trolling (Golf-Papez and Veer 2017) and cyberbullying (Quandt 2018). The spreading of fake news, disinformation, and conspiracy theories in UGC are forms of deviance as well (Frischlich et al. 2019).

Hence professionals face the challenge to handle a vast amount of data in tandem with their other daily tasks (Boberg et al. 2018). In order to ensure the website's quality, manual or automated moderation methods are used to check UGC before or after publishing (Hille and Bakker 2014; Singer et al. 2011; Veglis 2014). Moreover, media organizations have started to introduce new newsroom roles focused on navigating audience data and making sense of audience behavior. Engagement editors, social media editors, and analytics editors are expected to be more proactive, making sense of quantitative users' feedback to be able to predict their preferences (de-Lima-Santos and Mesquita 2021).

During the past few years, news media outlets have also started using artificial intelligence technology in new ways, from speeding up research to accumulating and cross-referencing data (Underwood 2019). The Huffington Post, for instance, has utilized a big data analysis for authenticating user comments. According to Stone (2014), the media organization conducted the statistical technique of conjoint analysis in order to determine the quality of comments coming from an anonymous person or those who have identified themselves either by name or by an avatar and from specific geographies.

6. Data Mining

While it is clearly evident that there are multitudes of potential applications for big data in the media industry, the fact remains that these datasets are so large and complex that in practice they are particularly unwieldy (Qiu et al. 2016). Even though the average journalist today possesses a much better set of ICT skills compared to a few years ago and is overall more digitally literate (Flew et al. 2012), big data as a phenomena are still, by their very nature, hard to access and work with. Stakeholders in the media industry that want to

utilize big data need to resort to special techniques and software capable of deciphering such large piles of information and breaking them down into a more usable and digestible format if they want to benefit from them (Wu et al. 2014). To that end, processes like data mining emerged as the solution to meet the needs of journalists in the media sphere.

Data mining is defined as a logical procedure used in order to search through very big amounts of data, with the purpose of discovering new, non-trivial information, which can subsequently be used to arrive at previously unknown conclusions (Ramageri 2010). By this definition, it is immediately obvious that data mining and big data go hand-in-hand when it comes to journalistic practices. As stated before, the nature of big data renders them inaccessible to being processed by humans, or even by simple software, because of various factors that make them hard to understand and compute. To that end, data mining is often utilized by journalists to overcome those obstacles, as it can enable the use of big data in order to uncover new and interesting connections between variables, which can lead to the discovery of crucial information for framing, or even creating a news story from scratch (Latar 2015). This is often accomplished by highlighting specific patterns within these huge amounts of data that can lead to interesting conclusions (Ramageri 2010). In other words, data mining tools enact the role of a mediator between journalists and big data, as they allow for the retrieval of potentially useful information from large datasets that would otherwise be inaccessible (Kotenidis and Veglis 2021).

The correct utilization of those tools has introduced myriads of possibilities for many stakeholders in the media industry, as it has paved the way for big data journalism, which, not only transformed the process of reporting, but also impacted the very format of the news story itself (Carlson 2015). Journalism managed to incorporate big data into its practices in a way that influenced the internal logic of the profession (Tandoc and Oh 2017), as the availability of these large datasets has introduced a more data-driven approach for many workers in the field.

At the same time, however, the changes brought about by the implementation of these new technologies have also turned the Media sector into a very volatile industry, when it comes to the skills required to succeed in it (Hammond 2017). This is due to the fact that workers are now forced to acclimate themselves with many new tools and techniques which were not necessary in the past, bringing the accessibility of data mining and other similar advanced journalistic methods into question. Simply put, the average journalist today needs to be proficient in many more technological fields than in the past in order to meaningfully compete within the confines of the media industry, a fact that has not gone unnoticed by researchers (Carlson 2015).

Similarly, the utilization of big data in journalism also comes with its own set of ethical considerations, since the data-driven approach it encourages does not accommodate for some of the traditional journalistic values like minimizing harm from uncontrollable information dissemination. According to Fairfield and Shtein (2014), this technological shift caused by big data highlights the need for the formation of a new ethical paradigm, not only for journalism, but perhaps for social sciences at large, with a potential re-defining of the relationships between researcher and research subject. As the authors point out, technological progress in journalism is meant to shift costs and create new opportunities. Based on that, this new paradigm hinges mostly on the ability of big data to provide on the first front, without compromising the second one. The new possibilities afforded by big data should extend to both the workers in the media sphere, as well as the audience, which should not be treated as just a source of data, but rather as a participating actor inside this ever-growing landscape of information, studied under an ethical research context.

Overall, however, the footprint of data mining in the field of journalism has clearly been a positive one. The ability to retrieve information from big data and the correct integration of this new innovation was able to expand the reaches of traditional journalism and introduce many new elements into the profession, which serve as the logical next step during an age of rapid technological growth.

7. Future Developments

The next version of today's Web is called the Semantic Web (SW). According to Tim Berners–Lee's article in the journal Scientific American, the SW is an extension of the current Web, in which information is given well-defined meaning, better-enabling computers and people to work in cooperation (Berners-Lee et al. 2001). It is anticipated that the SW will address the current Web's lack of structure by linking information from disparate sources and systems to ensure a more easy-to-use, efficient, and valuable scenario. When the SW will be fully developed, we will be talking about a Web of data, where every piece of information will be accompanied by its semantics, and its relations with the others will be fully clarified (Choudhury 2014). Hence, we will experience an interconnection of concepts, rather than just documents. In such a Web environment, both humans and machines will be able to work with vast amounts of information across several domains (Necula 2020) more meaningfully.

The technologies on which the SW is based are called Semantic Technologies (STs), and they refer to a set of programming languages and standards with common exchange protocols and data formats (Coronado et al. 2015). As Antoniou et al. (2012) point out, SW is an Internet service with advanced technological features. Their development is considered crucial for the integration of the SW, as these features appear to be able to overcome many information management obstacles of today's Web. Although SW constitutes nowadays mostly a research topic in Academia, its technologies are gradually becoming trends among different types of businesses (i.e., e-commerce), because of the facilities they provide (Necula et al. 2018; Rhayem et al. 2020). In short, the exploitation of ST provides to the content of the Web a commonly accepted structure, an explicit and comprehensible description and consensus that facilitates dataset sharing, querying, reuse, and integration (Rhayem et al. 2020). Their overall purpose is to provide a means for computers to understand data from a human point of view and process them accordingly (Adedugbe et al. 2020).

In this context, SW implies a type of Web that interprets searchable content and thus delivers appropriate and relevant information according to a fine understanding of the needs of users (Yen et al. 2015). Such a technological advancement will definitely unlock various possibilities of data exploitation for journalists and media professionals. However, today, we are witnessing the following paradox: On the one hand, we see journalists and media professionals welcome the wealth of big data in the current Web due to its dynamics and potential exploitation benefits, but on the other hand, we see them worry because they cannot make the most out of it due to the lack of an efficient infrastructure where accurate search, quick discovery, easy acquisition, and in-depth analysis can be realized. Getting the correct information at the right time is one of the prime demands of the digitization process (Bartussek et al. 2018) for every professional sector. Subsequently, an organization's or a journalist's competitive edge depends on asking complex questions across distributed data (Stardog 2019) and getting valid answers. Given the features of today's published data as web-based information, numeric, totally unstructured and only in some cases semi-structured, certain related tasks can become difficult to complete.

Hence, the challenge here is to derive human-conceivable meanings from big loads of machine-readable data, in order to draw valid conclusions. This issue can be addressed by examining the convergence of the SW and big data (Bello-Orgaz et al. 2016; Ahmed and Ahmed 2018). While we are at an early stage (Gross 2014), it seems that exploitation of big data through the use of ST is a real-life scenario. These technologies are able to ease all processes that are performed in the media, such as self-cataloging, enrichment, research, and content retrieval by using semantic-oriented features, such as diffuse searches, natural language queries, or multilingual searching (França et al. 2021). The ST (Fernàndez et al. 2018) along with a number of big data tools (Mujawar and Kulkarni 2015) are laying the groundwork to provide practical solutions to day-to-day tasks (i.e., documentation, writing) of both journalists and media organizations (Horrocks et al. 2016). Overall, the prospect of transforming big data, a pile of unstructured information that impractical to

use, into a structured dataset of great value seems to be a sign of the digital transformation of organizations (Lippell 2016).

Maybe this is why we see many leading media organizations adopting low-level ST, to interlink and connect information, among others. The BBC (British Broadcasting Co) (Raimond et al. 2010), New York Times, Thomson Reuters (Curry et al. 2010), NRK (Norwegian National Broadcaster) (Engels et al. 2007), Agence France Press, AP (Associated Press) (Underwood 2019), and several Danish news medias (Ildor 2020) are the ones who have already realized that ST is the key to the future. SW's rich environment, promising philosophy, various services, advanced technologies, and numerous applications are considered to be the next step toward a contemporary Web landscape.

8. Conclusions

This paper attempted to explore the utilization of big data by media organizations. Various areas where big data can be used have been reviewed, namely analytics, consumption, investigation, collaboration, and data journalism. The number of areas where big data can be applied today is continuously growing (Newman et al. 2018). Big data are playing and will continue to play a very important role in the success of media organizations (Veglis and Maniou 2018). They are a useful tool that allow media organizations to monitor users' reactions toward their products, facilitate user involvement in news production, create interesting news that is based on data, and offer users valuable information that is hidden in the data and would be out of reach for them. Big data utilization can facilitate media organizations to impose several levels of interactions between them and their audience, thus strengthening their relationship and building trust and loyalty, which are the most important values in the media sector. The latter is very important in today's post-truth era, where media organizations receive a lot of criticism for the issue of fake news and experience considerable competition from citizen journalists and legacy internet organizations (Facebook, Google, etc.) that have direct access to billions of internet users (Katsaounidou et al. 2018).

All of the above is precisely the reason why, amidst this newly developed landscape, it is important to consider the ethical aspects of big data utilization when it comes to journalism and social sciences in general. As explained throughout this manuscript, the versatility of big data allows for many new possibilities in a variety of sectors in media, but that simultaneously means that they have other far-reaching ethical implications as well. As with any new technological development which seeks to fundamentally change how the journalistic profession is being conducted, like the introduction of algorithmic technology for example, big data also have their fair share of ethical considerations. For instance, their utilization in storytelling and journalistic research could prove problematic in regard to the responsibilities of the journalist towards their audience. The consideration of the rights of an individual is a staple in social sciences, from which journalism borrows a lot in terms of ethics, but when it comes to big data, there are no "individuals" but rather millions of cases that form a larger interconnected sum (Lewis and Westlund 2015). As a result of this, it is impossible to accommodate some of the traditional journalistic practices, like asking for informed consent from the participants of a story or a study. This is particularly relevant when it comes to big data utilization on behalf of media organizations, as this data is routinely used for the creation of algorithmic solutions that automate various procedures in the content consumption and dissemination space, further feeding into the issue of algorithmic transparency and the role of these new innovative technologies in the realm of privacy (Diakopoulos 2015). These aspects are often overlooked when it comes to big data utilization in journalism, as the appeal of the benefits they provide can sometimes overshadow the minute details of how all that information is being collected or disseminated in the first place. This is the reason why it is worth examining the ethical consequences of big data in journalism more closely moving forward, especially when it comes to the average news consumer who could be disproportionally affected by any missteps in this regard.

In this diverse and continuously changing environment, media organizations have now the chance to employ semantic technologies in order to achieve better exploitation of big data. As was described in the previous section, such technologies have already started to attract the attention of big media, and it is expected that soon, they will start to be adopted by medium-level companies. As our world today is rapidly transforming into a data-driven society, it is quite obvious that data (and especially big data) will continue to play a very important role. This is something that media organizations cannot ignore, so the exploitation of big data is considered to be a necessity. The latter can also help them prepare for the new development in the narrative which will be based on text (chatbots) and speech (home assistants, i.e., Amazon Echo/Alexa, Google home/assistant) conversations (Veglis and Maniou 2019; Veglis and Maniou 2018). The future extension of this work will focus on the evaluation of big data exploitation in the examined areas of interest by media companies.

Author Contributions: Conceptualization, A.V.; investigation, T.S., K.P., C.K. and E.K.; writing—original draft preparation, A.V., T.S., K.P., C.K. and E.K.; writing—review and editing, A.V., T.S., K.P., C.K. and E.K.; visualization, K.P.; supervision, A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

Adedugbe, Oluwasegun, Elhadj Benkhelifa, Russell Campion, Feras Al-Obeidat, Anoud Bani Hani, and Uchitha Jayawickrama. 2020. Leveraging cloud computing for the semantic web: Review and trends. *Soft Computing* 24: 5999–6014. [CrossRef]

Ahmed, Jeelani, and Muqeem Ahmed. 2018. Big data and semantic web, challenges and opportunities a survey. *International Journal of Engineering & Technology* 7: 631–33.

Antoniou, Grigoris, Paul Growth, Frank van Harmelen, and Rinke Hoekstra. 2012. *A Semantic Web Primer*, 3rd ed. Cambridge: The MIT Press.

Appelgren, Ester, and Gunnar Nygren. 2014. Data Journalism in Sweden: Introducing new methods and genres of journalism into "old" organizations. *Digital Journalism* 2: 394–405. [CrossRef]

Bartussek, Wolfram, Hermann Bense, Thomas Hoppe, Bernhard G. Humm, Anatol Reibold, Ulrich Schade, Melanie Siegel, and Paul Walsh. 2018. Introduction to Semantic Applications. In *Semantic Applications—Methodology, Technology, Corporate Use*. Edited by Thomas Hoppe, Bernhard Humm and Anatol Reibold. Berlin: Springer, chp. 1. pp. 1–12. [CrossRef]

Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. 2016. Social big data: Recent achievements and new challenges. *Information Fusion* 28: 45–59. [CrossRef]

Bentley, Frank, Katie Quehl, Jordan Wirfs-Brock, and Melissa Bica. 2019. Understanding online news behaviors. In *CHI Conference on Human Factors in Computing Systems Proceedings*. Glasgow and New York: ACM, pp. 1–11. [CrossRef]

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American* 284: 34–43. [CrossRef]

Boberg, Svenja, Tim Schatto-Eckrodt, Lena Frischlich, and Thorsten Quandt. 2018. The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. *Media and Communication* 6: 58–69. [CrossRef]

Bodó, Balázs, Natali Helberger, Sarah Eskens, and Judith Möller. 2019. Interested in diversity. *Digital Journalism* 7: 206–29. [CrossRef]

Bradshaw, Paul. 2010. How to Be a Data Journalist. *The Guardian Data Journalism*. Available online: http://www.Guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide (accessed on 5 September 2022).

Carlson, Matt. 2015. The Robotic Reporter. *Digital Journalism* 3: 416–31. [CrossRef]

Charbonneaux, Juliette, and Pergia Gkouskou-Giannakou. 2015. "Data journalism", an investigation practice? A glance at the German and Greek cases. *Brazilian Journalism Research* 2: 244–67. [CrossRef]

Choudhury, Nupur. 2014. World Wide Web and Its Journey from Web 1.0 to Web 4.0. *International Journal of Computer Science and Information Technologies (IJCSIT)* 5: 8096–100.

Coronado, Miguel, Carlos A. Iglesias, and Emilio Serrano. 2015. Modelling rules for automating the Evented WEb by semantic technologies. *Expert Systems with Applications* 42: 7979–90. [CrossRef]

Cubbit, Sean. 2015. Data visualization and the subject of political aesthetics. In *Postdigital Aesthetics: Art, Computation and Design*. Edited by David Berry and Michael Dieter. London: Palgrave MacMillan, pp. 179–90.

Curry, Edward, Andre Freitas, and Sean O'Riáin. 2010. The Role of Community-Driven Data Curation for Enterprises. In *Linking Enterprise Data*. Edited by David Wood. Boston: Springer, pp. 25–47. [CrossRef]
de-Lima-Santos, Mathias-Felipe, and Lucia Mesquita. 2021. Data journalism in favela: Made by, for, and about forgotten and marginalized communities. *Journalism Practice* 1–19. [CrossRef]
De Maeyer, Juliette, Manon Libert, David Domingo, François Heinderyckx, and Florence Le Cam. 2015. Waiting for Data Journalism. *Digital Journalism* 3: 432–46. [CrossRef]
Deuze, Mark. 2006. Participation, remediation, bricolage: Considering principal components of a digital culture. *The Information Society: An International Journal* 22: 63–75. [CrossRef]
Diakopoulos, Nicholas. 2015. Algorithmic Accountability. *Digital Journalism* 3: 398–415. [CrossRef]
Engels, Robert, ESIS, and Jon Roar Tønnesen. 2007. A Digital Music Archive (DMA) for the Norwegian National Broadcaster (NRK) Using Semantic. *Semantic Web Use Cases and Case Studies*. Available online: https://www.w3.org/2001/sw/sweo/public/UseCases/NRK/ (accessed on 5 September 2022).
Evens, Tom, and Kristin Van Damme. 2016. Consumers' Willingness to Share Personal Data: Implications for Newspapers' Business Models. *International Journal on Media Management* 18: 25–41. [CrossRef]
Fairfield, Joshua, and Hannah Shtein. 2014. Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. *Journal of Mass Media Ethics* 29: 38–51. [CrossRef]
Fernàndez, Dèlia, Elisenda Bou Balust, Xavier Giró Nieto, Juan Carlos Riviero, Joan Espadaler, David Rodriguez, Aleix Colom Serra, Joan Marco Rimmerk, David Varas, Issey Massuda, and et al. 2018. Linking Media: Adopting Semantic Technologies for multimodal media connection. In Paper presented at ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks: Proceedings of the ISWC 2018 Posters & Demonstrations, Industry and Blue Sky Ideas Tracks co-located with 17th International Semantic Web Conference (ISWC 2018), Monterey, CA, USA, October 8–12, pp. 1–2. Available online: http://hdl.handle.net/2117/132968 (accessed on 5 September 2022).
Flew, Terry, Christina Spurgeon, Anna Daniel, and Adam Swift. 2012. The Promise of Computational Journalism. *Journalism Practice* 6: 157–71. [CrossRef]
França, Reinaldo Padilha, Ana Carolina Borges Monteiro, Rangel Arthur, and Yuzo Iano. 2021. An Overview and Technological Background of Semantic Technologies. In *Advanced Concepts, Methods, and Applications in Semantic Computing*. Edited by Olawande Daramola and Thomas Moser. Hershey: IGI Global, pp. 1–21. [CrossRef]
Frischlich, Lena, Svenja Boberg, and Thorsten Quandt. 2019. Comment sections as targets of dark participation? Journalists' evaluation and moderation of deviant user comments. *Journalism Studies* 20: 2014–33. [CrossRef]
Golf-Papez, Maja, and Ekant Veer. 2017. Don't feed the trolling: Rethinking how online trolling is being defined and combated. *Journal of Marketing Management* 33: 1336–54. [CrossRef]
Gross, David. 2014. Big Data and the Semantic Web—What Does It Really Mean? *Express*. Available online: https://www.express.co.uk/life-style/science-technology/529574/big-data-semantic-web-what-it-means (accessed on 5 September 2022).
Gruszynski Sanseverino, Gabriela, and Mathias Felipe De Lima Santos. 2021. Experimenting with user-generated content in journalistic practices: Adopting a user-centric storytelling approach during the covid-19 pandemic coverage in Latin America. *Brazilian Journalism Research* 17: 244–79. [CrossRef]
Hahn, Oliver, and Florian Stalph. 2016. Data Journalism in International Reporting. An Exploratory Study on Data-Driven Investigation of Foreign News Stories. Paper presented at the 2016 WJEC, 4th World Journalism Education Congress, Auckland, New Zealand, July 14–16.
Hammond, Philip. 2017. From computer-assisted to data-driven: Journalism and Big Data. *Journalism* 18: 408–24. [CrossRef]
Hermida, Alfred, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. Share, like, recommend. *Journalism Studies* 13: 815–24. [CrossRef]
Hilbert, Martin. 2016. Big data for development: A review of promises and challenges. *Development Policy Review* 34: 135–74. [CrossRef]
Hille, Sanne, and Piet Bakker. 2014. Engaging the social news user. *Journalism Practice* 8: 563–72. [CrossRef]
Horrocks, Ian, Martin Giese, Evgeny Kharlamov, and Arild Waaler. 2016. Using semantic technology to tame the data variety challenge. *IEEE Internet Computing* 20: 62–66. [CrossRef]
Ildor, Astrid. 2020. Semantic Web Applications for Danish News Media. In *WEBIST*. pp. 269–76.
Kalatzi, Olga, Charalampos Bratsas, and Andreas Veglis. 2018. The principles, features and techniques of data journalism. *Studies in Media Communication* 6: 36–44. [CrossRef]
Karypidou, Christina, Charalampos Bratsas, and Andreas Veglis. 2019. Visualization and interactivity in data journalism projects. Paper presented at 5th Annual International Conference on Communication and Management (ICCM2019), Athens, Greece, April 15–18, Athens: Communication Institute of Greece.
Katsaounidou, Anastasia, Charalampos Dimoulas, and Andreas Veglis. 2018. *Cross-Media Authentication and Verification: Emerging Research and Opportunities*. Pennsylvania: IGI-Global.
Kelly, John. 2019. Television by the numbers: The challenges of audience measurement in the age of Big Data. *Convergence* 25: 113–32. [CrossRef]
Kitchin, Rob. 2014. Big data, new epistemologies and paradigm shifts. *Big Data & Society* 1: 1–12. [CrossRef]
Kitchin, Rob, and Gavin McArdle. 2016. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society* 3: 1–10. [CrossRef]

Knight, Megan. 2015. Data journalism in the UK: A preliminary analysis of form and content. *Journal of Media Practice* 16: 55–72. [CrossRef]
Kotenidis, Efthimios, and Andreas Veglis. 2021. Algorithmic Journalism—Current Applications and Future Perspectives. *Journalism and Media* 2: 244–57. [CrossRef]
Ksiazek, Thomas B., Limor Peer, and Andrew Zivic. 2015. Discussing the news. *Digital Journalism* 3: 850–70. [CrossRef]
Latar, Noam Lemelshtrich. 2015. The Robot Journalist in the Age of Social Physics: The End of Human Journalism? In *The New World of Transitioned Media: Digital Realignment and Industry Transformation*. Edited by G. Einav. Berlin: Springer International Publishing, pp. 65–80.
Lewis, Seth C., and Oscar Westlund. 2015. Big data and journalism: Epistemology, expertise, economics, and ethics. *Digital Journalism* 3: 447–66. [CrossRef]
Lippell, Helen. 2016. Big Data in the Media and Entertainment Sectors. In *New Horizons for a Data-Driven Economy*. Cham: Springer, pp. 245–59.
Manosevitch, Idit. 2011. User generated content in the Israeli online journalism landscape. *Israel Affairs* 17: 422–44. [CrossRef]
Mujawar, Sofiya, and Soha Kulkarni. 2015. Big data: Tools and applications. *International Journal of Computer Applications* 115: 7–11. [CrossRef]
Murray, Scott. 2017. *Interactive Data Visualization for the Web: An Introduction to Designing with D3*. Sebastopol: O'Reilly Media, Inc.
Necula, Sabina-Cristiana. 2020. Semantic Web Applications: Current Trends in Datasets, Tools and Technologies' Development for Linked Open Data. *Informatica Economica* 24: 72–84. [CrossRef]
Necula, Sabina-Cristiana, Vasile-Daniel Păvăloaia, Cătălin Strîmbei, and Octavian Dospinescu. 2018. Enhancement of e-commerce websites with semantic web technologies. *Sustainability* 10: 1955. [CrossRef]
Nelson, Jacob L., and James G. Webster. 2016. Audience Currencies in the Age of Big Data. *International Journal on Media Management* 18: 9–24. [CrossRef]
Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David Levy, and Rasmus Kleis Nielsen. 2018. *Reuters Institute Digital News Report 2018*. Oxford: Reuters Institute for the Study of Journalism.
Newman, Nic, Richard Fletcher, Anne Schulz, Simge Andi, Craig Robertson, and Rasmus Kleis Nielsen. 2021. *Reuters Institute Digital News Report*, 10th ed. Oxford: Reuters Institute for the Study of Journalism.
Palomo, Bella, Laura Teruel, and Elena Blanco-Castilla. 2019. Data journalism projects based on user-generated content. How La Nacion data transforms active audience into staff. *Digital Journalism* 7: 1270–88. [CrossRef]
Parasie, Sylvain. 2015. Data-driven revelation? Epistemological tensions in investigative journalism in the age of 'big data'. *Digital Journalism* 3: 364–80. [CrossRef]
Qiu, Junfei, Qihui Wu, Guoru Ding, Yuhua Xu, and Shuo Feng. 2016. A survey of machine learning for big data processing. *EURASIP Journal on Advances in Signal Processing* 2016: 67. [CrossRef]
Quandt, Thorsten. 2018. Dark participation. *Media and Communication* 6: 36–48. [CrossRef]
Raimond, Yves, Tom Scott, Silver Oliver, Patrick Sinclair, and Michael Smethurst. 2010. Use of Semantic Web technologies on the BBC Web Sites. In *Linking Enterprise Data*. Edited by Wood David. Boston: Springer, pp. 263–83. [CrossRef]
Ramageri, Bharati. 2010. Data mining techniques and applications. *Indian Journal of Computer Science and Engineering* 1: 301–5.
Rhayem, Ahlem, Mohamed Ben Ahmed Mhiri, and Faiez Gargouri. 2020. Semantic Web Technologies for the Internet of Things: Systematic Literature Review. *Internet of Things* 11: 100206. [CrossRef]
Rogers, Simon. 2014. Introduction to Data Journalism. *Simon Rogers-Data Journalism and Other Curiosities*. Available online: http://simonrogers.net/2014/05/25/introduction-to-data-journalism/ (accessed on 5 September 2022).
Sandoval-Martín, María Teresa, and Leonardo La-Rosa. 2018. Big Data as a differentiating sociocultural element of data journalism: The perception of data journalists and experts. *Communication and Society* 31: 193–209. [CrossRef]
Saridou, Theodora, and Andreas Veglis. 2016. Participatory journalism practices in newspapers' websites in Greece. *Journal of Greek Media & Culture* 2: 85–101.
Singer, Jane B., David Domingo, Ari Heinonen, Alfred Hermida, Steve Paulussen, Thorsten Quandt, Zvi Reich, and Marina Vujnovic. 2011. *Participatory Journalism. Guarding Open Gates at Online Newspapers*. West Sussex: Wiley-Blackwell.
Singh, Priya. 2017. How Big Data Analytics Is Changing—The Media and Entertainment Landscape. *Analytics India Magazine*. Available online: https://analyticsindiamag.com/big-data-analytics-changing-media-entertainment-landscape/ (accessed on 5 September 2022).
Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. Big data: Big gaps of knowledge in the field of internet. *International Journal of Internet Science* 7: 1–5.
Spayd, Liz. 2017. A 'Community' of One: The Times Gets Tailored. *New York Times*. Available online: https://www.nytimes.com/2017/03/18/public-editor/a-community-of-one-the-times-gets-tailored.html (accessed on 5 September 2022).
Spyridou, Lia-Paschalia. 2018. Analyzing the active audience: Reluctant, reactive, fearful, or lazy? Forms and motives of participation in mainstream journalism. *Journalism* 20: 827–47. [CrossRef]
Spyridou, Lia-Paschalia, Maria Matsiola, Andreas Veglis, George Kalliris, and Charalampos Dimoulas. 2013. Journalism in a state of flux: Changing journalistic practices in the Greek newsroom. *International Communication Gazette* 75: 76–98. [CrossRef]
Stardog. 2019. Graph Identified as Top Technology Trend for 2019. Available online: https://www.businesswire.com/news/home/20190307005668/en/ (accessed on 5 September 2022).
Stone, Martha. 2014. *Big Data for Media*. Report. Oxford: Reuters Institute for the Study of Journalism.

Tandoc, Edson C., Jr., and Soo-Kwang Oh. 2017. Small departures, big continuities? Norms, values, and routines in The Guardian's big data journalism. *Journalism Studies* 18: 997–1015. [CrossRef]
Thurman, Neil, Judith Moeller, Natali Helberger, and Damian Trilling. 2018. My Friends, Editors, Algorithms, and I. *Digital Journalism* 7: 447–69. [CrossRef]
Tong, Jingrong. 2015. Chinese journalists' views of user-generated content producers and journalism: A case study of the boundary work of journalism. *Asian Journal of Communication* 25: 600–16. [CrossRef]
Turnbull, Sue. 2020. Imagining the Audience. In *The Media & Communications in Australia*. Edited by Stuart Cunningham and Sue Turnbull. London: Routledge, pp. 59–72.
Underwood, Corinna. 2019. Automated Journalism—AI Applications at New York Times, Reuters, and Other Media Giants. Emerj—The AI Research and Advisory Company. Available online: https://emerj.com/ai-sector-overviews/automated-journalism-applications (accessed on 5 September 2022).
Veglis, Andreas. 2009. Cross media Communication in newspaper organizations. In Proceedings of the 4th Mediterranean Conference on Information Systems, Athens, Greece, September 25–27, p. 37.
Veglis, Andreas. 2012. Journalism and Cross Media Publishing: The case of Greece. In *Handbook of Online Journalism*. Edited by Eugenia Siapera and Andreas Veglis. Hoboken: Blackwell Publishing, pp. 209–23.
Veglis, Andreas. 2014. Moderation Techniques for Social Media Content. Paper presented at HCI International 2014, Crete, Greece, June 22–27, Cham: Springer, pp. 137–48.
Veglis, Andreas, and Andreas Pomportsis. 2014. Journalists in the age of ICTs: Work demands and educational needs. *Journalism & Mass Communication Educator* 69: 61–75.
Veglis, Andreas, and Charalampos Bratsas. 2017. Journalists in the age of data journalism: The case of Greece. *Journal of Applied Journalism & Media Studies* 6: 225–44.
Veglis, Andreas, and Theodora Maniou. 2018. The mediated data model of communication flow: Big data and Data Journalism. *KOME: An International Journal of Pure Communication Inquiry* 6: 32–43. [CrossRef]
Veglis, Andreas, and Theodora Maniou. 2019. Chatbots on the Rise: A new Narrative in Journalism. *Studies in Media & Communication* 7: 1–6. [CrossRef]
Venturini, Tommaso, Mathieu Jacomy, Liliana Bounegru, and Jonathan Gray. 2018. Visual network exploration for data journalists. In *The Routledge Handbook of Developments in Digital Journalism Studies*. London: Routledge, pp. 265–83.
Weaver, David H., and Lars Willnat. 2016. Changes in US journalism: How do journalists think about social media? *Journalism Practice* 10: 844–55. [CrossRef]
Wu, Xindong, Xingquan Zhu, Gong-Quing Wu, and Wei Ding. 2014. Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering* 26: 97–107. [CrossRef]
Yen, Neil Y., Chengcui Zhang, Agustinus Borgy Waluyo, and James J. Park. 2015. Social Media Services and Technologies Towards Web 3.0. *Multimedia Tools Application* 74: 5007–13. [CrossRef]
Young, Mary Lynn, Alfred Hermida, and Johanna Fulda. 2018. What makes for great data journalism? A content analysis of data journalism awards finalists 2012–15. *Journalism Practice* 12: 115–35. [CrossRef]