Book Review

BIG DATA JOURNALISM: WHAT'S THE BIG DEAL?

Majda TAFRA, RIT Croatia, majda.tafra@croatia.rit.edu

Journalism in an Era of Big Data: Cases, concepts, and critiques (Journalism Studies) Edited by Seth C. Lewis, (2017), Routledge, Taylor & Francis (Kindle edition)

This book offers the first step in understanding what big data means for journalism. It was originally published as a special issue of Digital Journalism. Seth C. Lewis, its editor, is the inaugural Shirley Papé Chair in Electronic and Emerging Media in the School of Journalism and Communication at the University of Oregon, Eugene, OR, USA. He is also a visiting fellow with the Information Society Project at Yale Law School, New Haven, CT, USA. His widely published research explores the digital transformation of journalism, with a focus on the human–technology interactions and media innovation processes associated with data, code, analytics, social media, and related phenomena.

Introducing the topic, Seth C. Lewis, who also authors the first article in the book, will explain to practicing journalists, often skeptical of the words *big data* and *journalism* in the same sentence as well as those enthusiasts who call themselves "data journalists" whatever the phrase meant, that regardless of attitudes on the subject of big data journalism, what is needed is a rational approach and scholarly scrutiny, but also, some critique to cool down all the celebrating promises of reinventing news through the potential of "big data." That is the context in which this book consisting of eight chapters by various authors operates with a goal of exploring a range of phenomena at the junction between journalism and the social, computer, and information sciences. It implies all known contexts: digital information technologies being used in a newsroom; the algorithms, the analytics, applications, and automation.

Not only sociologists, but first of all media experts will ask what are the implications of this phenomenon for journalism. If asked what is it they do, journalist would probably answer they ask questions in order to get the answers and let the public know the truth. So, we speak of a profession that has its own development, its norms, routines, ethics and operates through media and relevant organizations which also have their own management and development. Not even to mention the underlying epistemology of journalism, the commitment to produce knowledge, make it known to the public and help in making sense of it.

As this book is being reviewed in the midst of the war in Ukraine, where starting from semantics (*war*, *aggression*, *military operation* – terms used exclusively in connection with who is using them and to which side in the conflict one belongs), going to deliberate use of video footage or lack of it (again depending on the side of the conflict) and the overuse of leading personality labels and extensive use of public appearances, the question what do big data have to do with the truth anyway, seems to be logic. The guest editor has a different, even broader question: What is the big deal about big

data?

Indeed, from a point of view of communication studies, what is the big deal? Lewis claims there are legitimate reasons to ask this question. First of all, it is being asked all the time in communication academic and journalistic circles, in social sciences, media and specifically in journalism and in methodology of computational social sciences. The second reasons he mentions is related to whose interest and with what purpose are big data promoted as a solution to various social problems. And finally, somebody must admit loudly that what we are dealing here with is, as he calls it "indeterminate set of leading-edge activities and approaches". They might be innovative, bring the whole new light in the darkest corners of the room. Or not. Who is to tell?

This is the explanation of the title of this book - "Journalism in an era of big data". There are two reasons, he explains. This era is characterized by an overwhelming volume and variety of digital information which are produced for and by humans. Daily, just by living in a digitalized world people create their own trails of data which are seldom observed scientifically. His second reason is found in the major development and advances in computing processing, machine learning , algorithms, and data science, which all enables various organizations and researchers to analyze this, as he calls it "shadow" layer of public life. All this is particularly important to understand this intersection of technology media and society, where journalism is positioned. Not much literature exists about the role of data in journalism. The whole issue, as Lewis claims it, is only beginning to get more attention first targeting journalism professionals with industry-facing reports, algorithms, and various debates on "quantification of journalism". This special issue is an effort to outline the state of the research in emerging domain, so that we try to understand what is becoming of journalism.

The issue is seeing journalism as interpolated through the conceptual and methodological approaches of computation and quantification, between ideation of computational and mathematical mindset on one side in newsroom and being ready to deconstruct and critique those same mindsets.

In the text titled *Clarifying journalism `s Quantitative Turn, A typology for evaluating data journalism, computational journalism and computer -assisted reporting*, Mark Coddington, having outlined the concepts of the open source culture, data driven journalism practices, computer-assisted reporting (CAR), data journalism, computational journalism, classifies and differentiates them in a typology that examines four dimensions: two of them professional (professional expertise versus networked information) and transparency versus opacity. One, big data versus targeted sampling is epistemological, and the final one has a professional/moral dimension – the vision of an active versus passive public. The dimensions are intended to serve as ideal forms against which individual cases and genres might be compared. There is a lot of overlapping between dimensions which is why the typology is not meant to be a definitive placement of these genres, but just an initial guide used to evaluate any computational or data oriented project, tool or organization.

He concludes that his typology is only an initial attempt to classify more systematically these data-driven journalistic practices, but these dimensions are hardly the only ones differentiating them. Since this area of journalism remains unsettled, new dimensions and forms of practice may emerge over the next several years. Still, this typology indicates a significant gap between the professional and epistemological orientations of CAR, on the one hand, and open-data journalism and

computational journalism, on the other This divide, as he explains, has its origins in the cultural background from which each has approached journalism: "CAR arose out of an effort to marry social science with modern professional journalism, and especially investigative journalism. Data journalism and computational journalism, on the other hand, have arisen from the intersection of professional journalism with open-source culture." (p.24). He points out that in academic communication field we still have an audience-centric perspective and our one-dimensional understanding of data-driven journalism, and the public, could be extended with a new approach If a quantitative turn is indeed occurring within journalism, it is important we research and understand how it changes the alignment with the profession's traditional values, practices, and the public.

In the second chapter the text *Between the Unique and the Pattern*, C.W. Anderson focuses on historical tensions in our understanding of quantitative journalism. His proposal is to analyze the history of the relationship between journalism and big data arguing that data need to be looked at as a particularly material procedural substrate (interviews, documents, observations and other journalistic genres). That approach helps in dealing with tensions between story and data. Will this tension be meaningful in a dozen years or so, he asks? History make help in answering that question. At the beginning of this millennium, some twenty years ago, who could have thought of "audience" contributing to journalist production? Today, in this world of social media we talk of billions of potential journalists, and great number of those who in fact do contribute to journalist content. It was radical then, today it is a common practice. He argues that what is radical in journalism today "may be its very conservatism: the fact that it exists at all as a relatively professionalized cadre of public information producers whose agenda is not entirely determined by the wishes of the audience." So, ten years from now, the big data debate may seem irrelevant as we shall live in the world of various, of measurement, outcomes assessment, and maybe, as he foresees it "increasingly narrow news production tailored to the consumption of niche audiences". In such a world what might be important about journalism maybe will be how journalism embraces other forms of information, not necessarily quantitative. " These other forms of knowing are possible because they once existed, and thus, they can exist again." (p. 42).

Sylvain Parasie in *Data Driven Revelation?* focuses on epistemological tensions in particularly in investigative journalism in the age of big data with a purpose to contribute to the analysis of how technology affects the epistemologies of journalism. Contributions are in two streams of scholarship: the future of investigative reporting and the role of technology therein. The author extensively elaborates the point that that data-processing artifacts can be used to enhance the collective organization of an investigation. The second contribution is in the study of journalistic knowledge. Since news organizations now experience alternative ways of producing justified beliefs from data, the studies are needed to investigate how these new practices shape the way news organizations globally deal with the processing of vast amounts of data.

An interesting case is analyzed in that context by Mary Lynn Young and Alfred Hermida in a text *From Mr. and Mrs. Outlier to central tendencies focus on the case of Computational journalism and crime reporting at the Los Angeles Times* . They are dealing with the specific case of the so called *Homicide Report* to illustrate a central tendency in understanding innovation in computational journalism, which has implications for how we assess technology adaptation in legacy media institutions. They try to apply a more critical sociological approach to the emergence of computational journalism.

Their findings show an uneven process of technology adoption/adaptation that builds on changing crime news norms and practices. The *Homicide Report* began using blog technology to support a more systematic approach to crime journalism with a public health agenda. As they explained "its first iteration emerged out of contemporary changes in the definition of crime news and contained early traces of computational thinking" This then shaped subsequent innovation. As a results of a later, more explicit adoption of computational journalism thinking and techniques, a new class of journalist was hired with specific expertise, extending the professional to both non-human crime journalists, so, the focus of resources was shifted to computational approaches to crime journalism in Los Angelos Times which had already previously had a history with CAR. The results were new forms of *Homicide Report*, as an interactive database and map. Later, journalists focused on the competitive possibilities for systemic coverage, transparency, and audience engagement. As they point out in the conclusion, "it can be argued that the approach masked a paradoxical shift in the professional role of the crime journalist, while, at the same time, nurturing the emergence of new and powerful identities of computational journalist in both its human and non-human forms" (p. 74)

Nicholas Diakopoulos in *Algorithmic Accountability* deals with Journalistic investigation of computational power structures identifying algorithmic power as worthy of scrutiny by computational journalists interested in accountability reporting. He offers a basis for understanding algorithmic power in terms of the types of decisions algorithms make in prioritizing, classifying, associating, and filtering information. In addition, he presents five case studies, which contribute to "delineating algorithmic accountability methods in practice, including challenges and considerations about the variable observability of input–output relationships as well as identifying, sampling, and finding newsworthy stories about algorithms." (p. 91), The case studies show that reverse engineering the input–output relationship of an algorithm can elucidate significant aspects of algorithms such as censorship. The paper also discusses challenges to the further application of algorithmic accountability reporting and shows how transparency might be used to effectively adhere to journalistic norms in the use of newsroom algorithms.

Matt Carlson elevates the subject to the topic of *The Robotic Reporter focusing on automated journalism and the redefinition of labor, compositional forms, and journalistic authority.* He is dealing with the increased practice of automated news content creation, how it alters the working practices of journalists and how it affects larger understanding of what journalism is. On the positive side, he states that this frees up journalists to pursue fewer mechanical stories. They also get a helping system capable of finding patterns easily missed by human perception. On the negative, what it boils down to are increased layoffs, polarizing personalization, and the commoditization of news writing. Beyond the related empirical questions, he insists this is the time to formulate critical questions for future research. What is going to happen to journalism if automated journalism would play a central role in the news landscape? There will be constraints, though, even as the system gets smarter mainly in available data and narrative-creating abilities. Finally, will an increase in algorithmic judgment lead to a decline in the authority of human judgment. This is, he claims , perhaps "the central question at stake with the technological drama surrounding automated journalism".(p. 109).

Waiting for Data Journalism by Juliette De Maeyer, Manon Libert, David Domingo, François Heinderyckx, and Florence Le Cam is another case study, qualitative assessment of the anecdotal take-up of data journalism in French-speaking Belgium. The results presented reveal that news

organizations have different approaches to data journalism as do journalists, editors, and human resources coordinators. The definitions of data journalism were “slippery” with a tension between each part of the doublet, data and journalism. The history of data journalism in Belgium indicated a trend that plateaued at the stage of the early adopters who engage in the production of “ordinary” data journalism but without indication it would “evolve towards a wider adoption, let alone a mainstream practice”. They even speak of resignation on part of those engaged in the concrete practice of data journalism which replaced a brief euphoric phase at the beginning. Talking to journalists revealed a number of obstacles, many of which are specific to data journalism like lack of both data and resources which they also put in the context of small market discursively populated by many people, tools, and organizations.(p. 124). They point out that the arrival of data journalism should also be seen in the larger context of other instances of new or evolving professional practices (e.g., multimedia, engagement with the audience) associated with the adoption of networked digital technologies in the newsrooms.

The book finishes by a concluding article by Seth C. Lewis and Oscar Westlund titled *Big data and journalism,* which deals with epistemology, expertise, economics, and ethics . Their various conceptual lenses - epistemology, expertise, economics, and ethics enable them to systematically explore both contemporary and potential applications of big data for the professional logic and industrial production of journalism. In the conclusion, thy point out that journalists and news organizations are seeking to make sense of, act upon, and derive value from big data during a time of exploration in algorithms, computation, and quantification and that developments of big data potentially have great meaning for journalism’s ways of knowing (epistemology) and doing (expertise), as well as its negotiation of value (economics) and values (ethics). They argue that these approaches are but “starting points for undertaking future research on big data and the opportunities and challenges that it poses for journalism, media, and society.” A lot of work in investigating the turbulent area of big data in what is still known as journalism, the meaning of which is apparently being very much changed and will continue to change for many reasons, one of the main ones being the rise of big data.

So, what is the big deal with big data journalism?

If you are a journalist, big deal, indeed!