2025 IEEE International Professional Communication Conference (ProComm) | 979-8-3315-3061-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/PROCOMM64814.2025.00014

A Theoretical Perspective on the Issue of Standardization of Fact-Checking in Data Journalism

Almabek Ibraiymov
Maqsut Narikbayev University
ibraiymov.almabek@gmail.com

Abstract - Investigative data journalism relies on data as a source and a tool for uncovering a story. The principles of working with data, searching, verifying, analyzing and data-storytelling differ significantly from traditional journalistic practices, but at the same time are an organic part of the journalistic material and the established standards and ethics of journalism. Fact-checking in data journalism is often reduced to mathematical verification of the correctness of data processing, with little attention paid to many other factors affecting the correctness and ethics of the text: the origin of the data, the methodology and motivation for their collection, the correctness of interpretation, the contextualization of the analysis results, and the correctness of data presentation in visualizations. At the same time, fact-checking in the field of data journalism is also seen only as a technological process of reconciling calculations, and not as a unified and systematic verification of journalistic material at all its levels, from technical to ethical. In this article, the author tried to fill the gap between the academic field and the practices of data fact-checking developed in newsrooms: existing methods of fact-checking in different editions were investigated, and existing gaps in data fact-checking were described.

Index Terms - Data journalism, debunking, disinformation, fact-checking, fake News

INTRODUCTION

One of the foundations of the journalistic profession is the rigorous and thorough verification of the facts reported in a story. Many newsrooms have a practice of verifying any story before publication, either by a special unit or by other journalists who were not involved in that story, or through separate organizations specializing in fact-checking. Existing methods and ethics of fact-checking have evolved over the last century, and professional standards for the quality of journalistic materials have been developed on their basis.

Researchers at the Rutgers School of Communication and Information emphasize the following quality standards for journalism: media should provide comprehensive and relevant information in society and the world, information should be objective (accurate, honest, sufficiently complete, consistent with reality, verifiable, facts should be distinct from opinions), and unbiased (all sides and interpretations should be presented equally and unbiased) [1].

The British Broadcasting Corporation (BBC) sets such ethical standards [2]: freedom of expression, independence, adherence to the public interest, impartiality and objectivity, credibility, editorial ethics, protection of vulnerable groups, and protection of privacy. With the emergence in the modern technological environment of a fundamentally different source - data - fact-checking practices in general have not changed or adapted to the new information reality.

Data journalism has emerged as a separate field of journalism in the last decade: it is based on working with data, acquiring, processing, analyzing, presenting it in visual form and narrative based on the results of this work with data. In this case, data serves as a tool for uncovering certain stories, as proof of facts, and as a source of investigative journalism. This highly technical area of journalism requires knowledge of data science, data visualization, and often programming. Data journalism owes its emergence to the worldwide movement for open information, the availability and openness of data, and the technologies that make it possible to store and process large amounts of data.

LITERATURE REVIEW

The tools and methods of data journalism are increasingly used in investigative, news and explanatory journalism as one of the components of storytelling; basic skills of data analysis and visualization are becoming not just in demand, but a mandatory set of skills for a modern journalist. At the same time, even the world's major editorial offices have not developed a standard of fact-

2158-1002/25/$31.00 ©2025 IEEE
DOI 10.1109/ProComm64814.2025.00014

checking for data investigations, and there are no established global practices in this area; moreover, public discussion of the need for a different approach to fact-checking in terms of data analysis, interpretation, and visualization is conducted only in a very narrow circle of specialists. Even the International Network of Investigative Journalists in its recommendations on the study and use of data journalism does not mention fact-checking of data investigations as one of the stages of a journalist's work [3], and numerous fact-checking services, including automated ones, are aimed at working with "traditional" resources and do not consider data-based investigations.

Data handling techniques in the context of fact-checking are also often considered as a component of computational fact checking [4] for online information in its broadest sense, with fact-checking in data journalism often reduced to the correctness of technological processes of data collection and analysis.

From the author's point of view, in the era of fast reading, mass and rapid digitalization of information, availability and dissemination of data, fact-checking in data-materials becomes an extremely urgent problem both from the point of view of technological data analysis and from the point of view of correct interpretation and contextualization of the results of data-based research. Fact-checking of this kind requires special skills and is often inaccessible due to objective (lack of knowledge, technological skills, technical resources, etc.) and subjective (unwillingness, lack of time, trust in the author, low level of critical thinking, etc.) reasons, which allows both errors in data-materials and deliberate manipulation of data to achieve certain goals. The development of standards for fact-checking in data journalism and their open publication, as it seems to the author, would help to significantly improve the transparency and quality of data investigations. Most data-driven journalism is published in a digital environment and remains openly available for users to view, verify and validate, and is shared via social media and private messaging. This "long-form" and accessible nature of data-driven investigations implies a high degree of factual accuracy and transparency in data handling methods, as well as a certain level of expertise in the topics under investigation.

Researchers David Cheruiyot and Raul Ferrer-Conill [5] note that even as data becomes a significant source for media and other organizations that apply the tools of journalism in their work, the established practices of the West are most actively considered in the academic field, and research on data factchecking in the media is virtually non-existent. They argue that non-journalism organizations, such as international NGOs engaged in data-driven fact-checking, are forcing a revision and renewal of journalistic discourse regarding data-driven practices. Cheruiyot and Ferrer-Conill suggest that epistemologically, journalism is changing from evaluating facts provided by others and creating one's own knowledge to forms of knowledge and knowledge production that are closely linked to the audience's acceptance of that knowledge. In their view, this change is being driven by journalism's systems of quantification and data-driven practices. In the logic of fact-checking, data can become a journalistic source, which, within the existing discourse of objectivity, means a qualitative dimension of journalism. Researchers talk about the value of data as an explicit source of knowledge for its subsequent reproduction and dissemination, but do not dwell on factchecking practices directly in the field of data journalism, focusing on algorithmic factchecking systems.

Mark Coddington [6] points out that data journalism is based on the four pillars of openness - transparency, interactivity, changeability and complicity - and that the pervasiveness of data is changing established journalistic practices. He considers the main epistemological difference between data journalism and other digital methods in journalism to be the practice of complicity, in which journalists see the audience as co-authors in the search for truth and moral rights. Coddington also notes that visualization has become a distinct value in data journalism, one of the key features of the genre, suggesting that design is linked to the values of the journalist. He also talks about the peculiarity of data journalism, in which data is subordinated to journalistic values and the narrative of the story, with data being researched using scientific methods, and the evaluation of interpretation and contextualization often left outside the knowledge of journalists in expert environments. At the same time, the researcher does not reflect on the change in editorial practices of factchecking due to the emergence of data as a significant source.

Researchers Kennedy, Webet The above leads the author of the article to the need to fill the gap between the established practice of data journalism in newsrooms and the academic field, where factchecking in this genre is rarely and only partially discussed, not as a unified system with its own peculiarities, and the epistemological "heredity" of methods of verification and validation of information in data journalism is not comprehended. In this paper, the author examines existing editorial practices and the integrity of their approach, and based on the analysis proposes the foundations of his own system of factchecking in data journalism.

The above leads the author of the article to the need to fill the gap between the established practice of data journalism in newsrooms and the academic field, where factchecking in this genre is rarely and only partially discussed, not as a unified system with its own peculiarities, and the epistemological "heredity" of methods of verification and validation of information in data journalism is not comprehended. In this paper, the author examines existing editorial practices and the integrity of their approach, and based on the analysis proposes the foundations of his own system of factchecking in data journalism.

Reuters in the Handbook of Journalism [8] does not mention data as a source or investigative method at all, even though it regularly produces data materials and complex investigations based on data collection, analysis and interpretation. Data as a source requiring verification is also omitted entirely in the Verification Handbook [9] of the European Journalism Center [10]. Many editorial offices that conduct dataset investigations indicate in their materials the sources of data and sometimes descriptions of datasets, but do not provide information about research methods and fact-checking in datasets: among them ProPublica, The New York Times, The Guardian.

Validity assessment practices in most media outlets are limited to the technological aspects of data collection, cleaning, and analysis. For example, the Associated Press, in its public document "News Values and Principles" [11] limits itself to a brief remark about data: "Data for stories and visual presentations should be checked for completeness and validity. Data should be evaluated in terms of collection methodology, sample size, collection time, and the availability of other data that can confirm or challenge it. Combining more than one dataset in a presentation should be done carefully and transparently. Avoid comparing percentages and percentage points for a small sample, including raw numbers needed for general understanding. We must clearly distinguish correlations from causation." The widely recognized Data Journalism Handbook [12] of the European Journalism Centre discusses in detail the tools of working with data and how to process them but also does not address the issue of fact-checking such studies, the authors do not dwell on the quality of interpreting the results of data analysis and putting them into context.

The International Fact-checking Network (IFCN - International Fact-checking Network) [13] of the Poynter Institute has developed its own code of principles [14] for verifying information in media; it includes five principles, each of which is described by five or six requirements. We have highlighted from this number only those obligations for the participants of this community, which, in the opinion of the author of this article, are applicable, including to data:

- The same fact-checking standards are applied to verify all facts regardless of the party presenting them.
- Conclusions are based on the evidence obtained, rather than evidence being selected to support conclusions.
- The possibility of independent fact-checking by readers - community members provide maximum information and references to sources sufficient to reproduce the results of the investigation, where this does not compromise personal security and privacy.
- Transparency of the methodology used in sampling, fact finding, fact processing and factchecking.

At the same time, data as a source is not considered in principle in the IFCN standards, so fact-checking of data analysis and presentation is not mentioned. Some of the principles and obligations are described broadly enough to be applied to data investigations, while the key issues in the proposed framework are verification of sources, transparency of methodology and the possibility of replicating the journalist's actions to obtain the same results. The issues of data quality, correctness of its interpretation, and the correspondence of data to reality are not covered.

Paul Bradshaw, one of the pioneers of data journalism and a lecturer in journalism at universities in Birmingham and London, believes that one of the key factors in assessing the quality of data material should also be its ethics [15]: the accuracy of data presentation depends on its correct contextualization. Depending on the context, data can be differently visualized, interpreted, presented or even not used or published at all. For example, it is impossible to use Russian state data on the number of orphans as a basis for journalistic investigation: such data are provided by state structures [16], but the specifics of collecting this data are such that, in fact, they are meaningless - there is no unified methodology of accounting for different regions, the recommendations for collecting this data are not followed by the owners of datasets, and it is virtually impossible to check the integrity and reliability. Such data, although considered open, give a distorted picture of reality, and therefore cannot be used in investigative journalism, neither from the point of view of professional standards of journalistic quality nor from the point of view of ethics. Data, for example, on the number of children with disabilities in Russia [17] are presented by the responsible agencies (the Ministry of Labor and Social Protection, the Department for the Disabled) in a fragmentary manner, which also significantly distorts the idea of the real state of affairs - in this case, the journalist must collect and analyze data from other agencies: the Ministry of Finance, the Pension Fund, the Ministry of Health, and the Ministry of Education, where other data on children with disabilities are presented: accrual of pensions, special conditions of education, etc. The collection of indirect data in this case gives a completer and more relevant picture, but requires a more careful study of the data, correct correlations of interagency data, as well as expert knowledge in this area, allowing us to give a correct assessment of faceless figures and see behind them a living human history and the essence of the existing social problem.

In the Russian edition of 'Novaya Gazeta' [18], the factchecking process depends on the type of material. In the newspaper's dataset department, the general rule is to find as many different datasets with data on the topic of interest as possible and compile the most complete dataset, correct possible inaccuracies (remove repetitions, gather different parts into a single dataset), and obtain information from subject matter experts about the quality and reliability of the sources.

The publication “Important Stories” [19] practices the factchecking model proposed by OCCRP [20]: the author of the story puts footnotes with evidence for each fact; the factchecker, another data journalist in the editorial office, checks them. In this case, fact-checking essentially happens twice: first by the author, who can see inaccuracies and correct them when placing references, and then by a fellow journalist. In the process of checking, the correctness of each fact and statement in the text is evaluated; if the fact-checker has questions, they are solved with the author of the text; if the author and the fact-checker do not come to a common opinion, a comment is left in the text for the editor. All sources, calculations, dates, names, titles, correspondence of quotations to the records of conversations with experts and heroes of publications are checked. Since the material is checked by a data-journalist, the probability of errors in calculations and methods of data analysis is reduced.

The publication “Project.” [21] approaches the verification of data-based materials differently: after editing, the text together with data, calculations and references to sources goes to the fact checker - another journalist of the editorial office, most often not specialized in data investigations, who must find evidence for each sentence in the text. If no such evidence is found, or if there are errors, the fact-checker asks for clarification from the author. In this case, the fact-checker looks for evidence independently, which allows them to reduce the level of bias and pay attention to details and other sources that the author may have overlooked.

Data journalist Winnie de Jong [22] proposes this checklist for checking data correctness:

1) Check dates, spelling and duplicates, drop-down values. Remember that statistical significance is not new. Make sure trends are presented diachronically. Make sure the data correlates with reality.
2) Keep a data diary as you work, where you describe the sequence of activities and the data activities themselves. You should be able to reproduce your calculations.
3) Describe your working methods - the reader should understand how the author discovered the story in the data.
4) Keep a file with footnotes: provide each fact with a number, and for each number collect information: how you know the fact, the source, the proof. Correct any errors found in this file.

Methodology

The proposed methodology also focuses more on technical accuracy in the collection and processing of data, although there is mention of the need to contextualize it, and the methodology does not consider the assessment of data quality and the correctness of its presentation as stages of fact-checking.

Unlike many other journalistic genres, data research should be reproducible: anyone repeating the same actions with the same set of data should come to the same conclusions that the journalist came to. This, in the author's opinion, is one of the key indicators of the quality of data-journalistic work. Most data journalism is based on open data that is available for analysis by anyone on the web. In addition, data science as an exact discipline assumes absolute reproducibility of the result. The combination of these factors with the underlying principle of journalistic credibility leads the author to conclude that any investigation made in the public interest based on open sources must be verifiable and credible, in which case the validity of the analysis will be confirmed by its reproducibility. The verifiability of the information provided also formed the basis of the International Fact-checking Network (IFCN) principles. Therefore, accurate step-by-step logging of all actions performed with data becomes a prerequisite both in the work of a data journalist and in fact-checking in the editorial office and by volunteers. Such a log also makes it possible to track the correctness of actions and use of different analysis methods at all stages of data collection, cleaning, analysis and other operations with data and to calculate errors in them both during self-checking and during fact-checking. Storing all versions of the data will allow you to quickly return to the point where an error or inaccuracy occurred and further analyze correctly.

The reliability and reputation of the data source should also be assessed during the fact-checking process. If the data source is questionable, including the motivation for collecting and reporting the data, it is imperative to obtain comments from an expert in the field of study: why the data were collected, what questions they can answer, what limitations and exceptions there may be in the data.

The methodology of the original data collection should also be examined and considered in the analysis, as it affects the methods of data processing and analysis. For example, from January 15, 2020, to the current day, the parameters for counting COVID-19 coronavirus cases in China have already changed seven times [3]. This means that visualizing a trend based on this data as the same would be obviously erroneous, even though the source - the official statistical office of a large country - may be considered reliable.

In the view of the author of this article, the accuracy of data interpretation should also be assessed during fact-checking procedures in data histories, as often the same data can be used to support two contradictory views. For example, data on the number of women in government can be interpreted as both 'the number of women in government has doubled' and 'women are underrepresented in government' if the number of women parliamentarians has changed from 20 to 40 over the course of a year, out of a total parliament of 250. The context in which the data exists, the methods and purposes of data collection play an important role in the story and should also be verified in fact-checking.

Results

None of the above-mentioned fact-checking practices offers a stage of quality assessment of visualizations - from the necessity of such visualizations to the correctness and appropriateness of their execution. Such verification will avoid intentional or accidental distortions in the presentation of data and, consequently, the formation of a misconception of the subject in the public. Such a frequent example of data presented incorrectly and distorting the idea of reality can be called bar charts in the news programs of the TV channel “Qazaqstan”, where visualizations do not offer a point of reference, but display only the difference of indicators, excluding or reducing the values codified in such a chart. The author does not propose to evaluate creative solutions in the field of visualization but insists on the use of objective metrics for evaluating the quality of visualization. Such metrics are suggested, for example, by A. Bogachev [24]: accuracy of data presentation, correctness of chart type selection, correctness of specifying units of calculation and reference points, correct display of ratios of values, correctness of legend writing, etc.

As a result of studying data-fact-checking practices in newsrooms, the author formulated several recommendations on verification in data journalism, which can be implemented in newsrooms both those specializing in data investigations and those who sometimes resort to the tools of data journalism.

1) Storing the original data with their sources, dates of accessing and downloading the datasets. It is also reasonable to keep screenshots of web pages from which files were downloaded, because not always data owners fulfill the requirements for open and accessible storage of source data and delete them for various reasons, from ignorance of the rules of handling open data to attempts to hide information.
2) In cases when data were collected by the editorial staff (surveys, scraping - collection of data of a given type from a site or sites by special programs, etc.), it is necessary to describe in detail the methodology and tools of data collection.
3) Keep a detailed log of data handling. There are several options available to newsrooms. One is to enable automatic logging of all actions in data processing programs such as Excel, Google Sheets, Open Refine, and others. Another option is the open-source Jupiter software, the Jupiter Notebook web application allows you to log program code, equations, visualizations, and text in a single document. Importantly, Jupyter Notebooks are both human-readable and machine-readable documents. Recently, the combination of these approaches has been increasingly used in industry, e.g., BuzzFeed and The New York Times.
4) It is necessary to store all versions of the data - this will allow during fact-checking to detect at what stage errors may have occurred, as well as to return to previous versions and ensure the integrity and safety of the data.
5) A detailed description of the data analysis methodology. It is necessary to store all versions of the data - this will allow during fact-checking to detect at what stage errors may have occurred, as well as to return to previous versions and ensure the integrity and safety of the data. For certain types of data provided in a uniform way, collected using a single methodology, the research methodologies should be standardized to eliminate bias in individual cases. This is true, for example, for regular studies of statistical indicators, sample reports, etc. For all studies requiring non-standardized approaches, the methodology should be described in detail.
6) Expert evaluation of the correctness of the interpretation of the result and its contextualization by a professional in the researched area, with the expert's comments retained.
7) Peer review involving data journalists who have not been involved in the material, with the possibility of saving comments and edits made because of such a review.
8) Evaluation of visualization quality: how correctly the visualization presents data, whether it reflects the result of the investigation, whether it does not distort ratios, whether it displays all necessary values, whether it indicates values correctly, etc. The author believes that the system for assessing the quality of data visualization should be developed separately and serve as a kind of checklist for journalists and designers when creating visualizations, infographics, and retinographies.

Conclusion

Parts of these practices already exist in major publications. Some media outlets, such as Buzzfeed, sometimes publish their Jupiter Notebooks on industry platforms - this openness of investigations should show the unbiased and objective nature of the publication, but at the same time the selectivity of notebook publishing raises valid ethical questions. Certain publications, for example, Novaya Gazeta, in their data investigations necessarily indicate the methodology of data processing and the limitations of the study - this is an important reputational element that allows researchers not affiliated with the editorial board to verify the conclusions obtained by analyzing the data, which is in line with IFCN and OCCRP recommendations. Citing sources has become a de facto standard in data journalism, while editorials that describe in such detail how to handle data in investigative journalism are a minority.

References

[1] S. Lacy and T. Rosenstiel, "Defining and measuring quality journalism." Rutgers School of Communication and Information, New Brunswick, NJ, 2015, Available: https://www.issuelab.org/resources/31212/31212.pdf

[2] BBC, "Editorial guidelines." [Online]. Available: https://www.bbc.co.uk/editorialguidelines/guidelines

[3] Global Investigative Journalism Network, "Introduction to investigative journalism: Data journalism," [Online]. Available: https://gijn.org/resource/introduction-investigative-journalism-data-journalism/

[4] S. Cazalens et al., "Computational fact checking: A content management perspective," *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 1930–1933, 2018, doi: 10.14778/3229863.3229880

[5] D. Cheruiyot and R. Ferrer-Conill, "Fact-checking Africa: Epistemologies, data, and the expansion of journalistic discourse," *Digit. J.*., vol. 6, no. 8, pp. 964–975, 2018.

[6] M. Coddington, "Clarifying journalism's quantitative turn: A typology for evaluating data journalism, computational journalism, and computer-assisted reporting," *Digit. Journal.*, vol. 3, no. 3, pp. 331–348, 2014.

[7] H. Kennedy et al., "Data visualization and transparency in the news," in *Data Visualization in Society*, M. Engebretsen and H. Kennedy, Eds. Amsterdam: Amsterdam University Press, 2020, pp. 169–185.

[8] Reuters, *Reuters Handbook of Journalism*. [Online]. Available: https://www.mediareform.org.uk/wp-content/uploads/2015/12/Reuters_Handbook_of_Journalism.pdf

[9] C. Siverman, Ed., *Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage*, European Journalism Centre. Available: https://verificationhandbook.com/downloads/verification.handbook.pdf

[10] European Journalism Centre. [Online]. Available: https://ejc.net/

[11] The Associated Press, "Statement of news values and principles." [Online]. Available: https://www.ap.org/about/news-values-and-principles/downloads/apnews-values-and-principles.pdf

[12] J. Gray and L. Bounegru, Eds. *The Data Journalism Handbook 2: Towards a Critical Data Practice*, European Journalism Centre and Google News Initiative. Available: https://datajournalism.com/read/handbook/two

[13] Poynter, "The international fact-checking network." [Online]. Available: https://www.poynter.org/ifcn/

[14] IFCN Code of Principles. [Online]. Available: https://ifcncodeofprinciples.poynter.org/

[15] P. Bradshaw, "Data journalism," in *Ethics for Digital Journalists: Emerging Best Practices*, L. Zion and D. Craig, Eds. London: Routledge, 2015, pp. 202–219.

[16] EMISS, "Number of orphans and children left without parental care in residential social service institutions." [Online]. Available: https://fedstat.ru/indicator/41601

[17] EMISS, "Number of persons first-time recognized as disabled in the category of 'disabled child'." [Online]. Available: https://fedstat.ru/indicator/41623

[18] Novaya Gazeta. [Online]. Available: https://novayagazeta.ru/

[19] Important Stories. [Online]. Available: https://www.istories.media/

[20] Organized Crime and Corruption Reporting Project. [Online]. Available: https://www.occrp.org/en

[21] Proekt. [Online]. Available: https://www.proekt.media

[22] W. De Jong, "How not to be wrong," *Global Investigative Journalism Network*. [Online]. Available: https://gijn.org/2017/05/25/how-not-to-be-wrong/

[23] C. Campbell and A Gunia, "China says it's beating Coronavirus. But can we believe its numbers?," *Time*. [Online]. Available: https://time.com/5813628/china-coronavirus-statistics-wuhan/

[24] A. A. Bogachev, *Charts That Convince Everyone*. Moscow: AST, 2020.

About the Author

Almabek Ibraiymov holds a master's degree in social science and serves as a Senior Lecturer at the International School of Journalism, Maqsut Narikbayev University, Kazakhstan. His research interests include fact-checking, Open Data, data journalism, new media.