ORIGINAL ARTICLE

DOI: 10.51480/1899-5101.15.3(32).3

Fighting COVID-19 with data: An analysis of data journalism projects submitted to Sigma Awards 2021

Liis Auväärt
ORCID: 0000-0002-9786-1696
University of Tartu

Abstract: The COVID-19 health crisis has been heavily reported on an international scale for several years. This has pushed news journalism in a datafied direction: reporters have learnt how to analyse and visualise the statistical effects of COVID-19 on various sectors of society. As a result, in 2021, the international Sigma Awards competition for data journalism highlighted coverage of the pandemic. Using content analysis with qualitative elements, this paper analyses the shortlisted works covering COVID-19 from the competition (n=73). It focuses on the data references made by the teams – sources, type of both reference and data used – showing statistics from official institutions to be the most used type of data. It also lists the main problems journalists had to face while working on their projects. Most often these problems fell into two categories: specific characteristics of the project, mostly 'time consuming', and issues with data.

Keywords: COVID-19, data journalism, data literacy, datafication, journalistic skill

INTRODUCTION

In early 2020, reports on the COVID-19 virus claimed their place as the top news throughout the world. In the context of journalistic skills, the pandemic highlighted the need to keep pace with worldwide datafication, which can be phrased as the "quantification of aspects of life previously experienced in qualitative, non-numeric forms" (Engebretsen et al., 2018, p. 1). To datafy a phenomenon is to put it in a quantified format, so that it can be tabulated and analysed (Mayer-Schönberger & Cukier, 2013) – health being a highly datafied field.

Coverage of the coronavirus has posed challenges to journalists everywhere, as careful news production has been required to guide people and reduce uncertainty, but also to create balance and avoid the increasing spread of health mis

– and disinformation (Casero-Ripolles, 2020; Catalan-Matamoros & Elías, 2020). Analysing and visualising the spread of the virus, daily death numbers, the progress of vaccination procedures, etc. became a daily task for media outlets. Phrases such as 'mathematical model', 'flatten the curve', 'peak infection' and various other mathematical terms appear in both media and public discourse (MacDonald, 2021; Stephan et al., 2021).

In 2021, the data journalistic challenge of COVID-19 was highlighted by the Sigma Awards – an annual large-scale international competition for data journalism projects, successor to the discontinued Data Journalism Awards. In brief, the Sigma Awards aims to empower and enlighten the global community of data journalists by offering a platform, on which they can compete and discuss their work. As for the 2021 competition, the organizers openly declared the goal to spotlight the pandemic.

Emphasising the volume of work done covering COVID-19, the competition was indeed dominated by projects tied to coronavirus. In total, 545 data journalism projects from 68 countries were submitted. Thereafter, a pre-jury shortlisted 140 of these projects with more than half – 73 projects, 52% – reporting on issues tied to COVID-19.

The literature notes that little has been written about journalists regarding data skills and choice of sources (Wihbey, 2017). This paper aims to look at the data sources used by the teams. As background variables, the location and size of the teams will be analysed, then the geographical focus and topic of the projects will be described and data sources – number and types of source as well as types of data – used in the project. Finally, the main problems the journalists had to face while working on their project, as described by themselves, will be reviewed, thus helping better understand the skillsets currently possessed by data journalists and their level of technical and data literacy.

LITERATURE REVIEW

A SHIFT TOWARDS DATA JOURNALISM: NEW SKILLS AND PRACTICES

Nowadays, newsrooms are saturated with data. On the one hand data have become a tool that, through audience metrics and advertising revenues, shapes media industry. On the other hand, data provide the possibility for countless journalistic projects, which combined with increased computing power will help us reach a new level of understanding about the world (Marconi, 2020). Datafication has caused not only hybrid roles such as programmer-journalist, journalist-developer and hacker-journalist have come into play (Parasie & Dagiral, 2013; Royal, 2012) but has also transformed news media editors from being

hybrid journalist-information officers (Marconi, 2020). Throughout this paper, phrases like 'data journalism' will be used to mark the process of either or both reporting on data and using data for reporting. In newsrooms, data journalists represent a combination of traditional journalistic values and the values of open-source culture (Widholm & Appelgren, 2020).

The claim that data-driven journalistic practices are gaining momentum in newsrooms all over the world is supported by the growing corpus of research literature. In recent years numerous studies have been published discussing on various aspects of data journalism. From Europe, for example there was a representative quantitative overview of data journalism activities in German newspapers and by public broadcasters, which concluded that data journalism in Germany is well established (75% of media outlets) and mostly performed by individuals or small teams (Beiler et al., 2020). A study in Italy, based on interviews with full-time data journalists, highlighted the need for better journalistic education but also shed light on strategies for how data journalists generate and collect their own data (Porlezza & Splendore, 2019). Research from Scotland, Northern Ireland and Wales, applies the prism of material, performative and reflexive conceptual lenses to data collected from interviews with data journalists and data editors (Borges-Rey, 2020). A quantitative analysis of news produced by specialised data desks in Sweden's public service organisations, concludes that in data journalism hard and soft news attributes often appear close together in hybrid forms (Widholm & Appelgren, 2020). These studies, among others, confirm growing journalistic interest in data – with data journalists emerging, data teams created and journalists and programmers teaming up to work on stories. They also address the issue of data journalism being largely determined by public datasets – most often as freely available data from statistical offices.

For a journalist, a professional storyteller, data hold power. Several studies point out that interpreting data in a news piece has a psychological effect because it helps the journalist emphasize the importance of the issue and improve the quality of the story (Wihbey, 2017). Quoting data signifies the story due to the culturally embedded belief that "measured knowledge, expressed in numbers, represents undebatable truth that cannot be argued with" (Van Witsen, 2020, p. 1061).

Focusing on stories supported by or hidden in data presents an opportunity for the media to remain relevant and even strengthen their role in informing the people. Understanding the business of news is connected to the feeling of 'value' people get from reporters (Lewis, 2020). Scholars have proposed that journalists may reconceive themselves as 'knowledge brokers', who illuminate the process of expert knowledge production to their audience (Nisbet & Fahy, 2015). Coverage of COVID-19 has presented the media with this role very clearly,

as vast amounts of pandemic-related data and issues need to be addressed and presented in a way that is easily understandable to the public. The pandemic switched journalism into crisis mode, but unlike other crises, this one had "a direct and immediate impact on journalists themselves, their work routines, economic and technological resources, media as institutions and the societal and political environment" (Quandt & Wahl-Jorgensen, 2022, p. 924).

Undoubtedly, COVID-19 has been a highly datafied issue exemplified by the use of a variety of data such as: (a) healthcare-related statistics describing the situation locally and worldwide and (b) medical expertise to explain the nature of the virus; (c) spatial data that show the spread of the pandemic and (d) census data determining the areas in which hospitals might fall into crisis due to the number of potential patients; (e) financial data on government expenses for masks and medicine and (f) measurements of the loss felt by sectors of the economy.

This is where the 'expert knowledge' referred to earlier truly shines. Gathering and cleaning the data, collating and visualising them requires several skills; therefore, from a data journalist's point of view, data literacy and at least some technical knowhow is vital.

DATA LITERACY AND JOURNALISTIC SOURCES

The OECD (2016) defines information processing skills as consisting of literacy, numeracy and problem-solving in a technologically saturated environment. In data journalism research, the term is defined in a similar way: a mixture of statistical, numerical and technical capacities – being data literate means "being able to access, analyse, use, interpret, manipulate and argue with data-sets in response to the ubiquity of (digital) data in different fields" (Gray et al., 2018:, p. 2).

This definition misses the word 'quickly'. On any given day, the environment of a newsroom dictates the need to interpret data quickly yet accurately, which was amplified when COVID-19 struck. Although health and science have been common topics in data journalism since data is routinely available in these domains (Cushion et al., 2016; Loosen et al., 2020), the volume of health-related expert data that came with COVID-19 was a journalistic challenge. There was, and still is, a lot of data being published about the pandemic – for example, it was noted that Web of Science alone published more than ten COVID-19 related papers per hour (Makri, 2021) and in 2020 alone, researchers are estimated to have published 100,000–200,000 papers related to the virus (Stoto et al., 2022). Tackling the issues at hand requires reporters to be skilled in understanding data and finding reliable expert sources.

Building a network of trustworthy data sources, a peer review so to speak, helps save time, too because reporters aim not only to be accurate, but they also want to break the news before their competitors (Makri, 2021). Research by Stalph

(2018) analysing European quality news websites indicates that data journalists mainly rely on pre-processed data drawn by domestic governmental bodies. The literature shows this notion to be common. In analysing the relationship between data owners and journalists, several researchers have remarked in the past decade that there is a strong dependence on data provided by government institutions and other non-commercial organisations (NGOs, research institutes, etc.) – data which are publicly available or available upon request (Beiler et al., 2020; Parasie & Dagiral, 2013; Van Witsen, 2020; Young et al., 2018).

One can expect that the pandemic further solidified government bodies, universities and other research institutes as data sources. For example, this has proven true in the case of the COVID-19 vaccination process in Spain (Catalan-Matamoros & Elías, 2020). Covering COVID-19 presented an issue so novel that journalists had to start explaining it to the public from scratch because "common knowledge" about it did not yet exist. Thus, reliance on public datasets and help from various expert sources to interpret them from various angles seems only logical. Furthermore, creating a dataset on your own might require time, skills, people and funding that are not available to all data journalists on a regular basis. It can be argued that this also shapes the variety of stories being published on the matter of the pandemic.

Although the pandemic was a grim subject, the COVID-19 global health crisis offers the chance to look at worldwide data journalism projects focusing on a mutual theme. In doing so, it provides an opportunity to better understand the current state of data journalism. An analysis of pandemic-related projects shortlisted by Sigma Awards proves useful in several aspects. First, the analysis gives an overview of what data projects by journalists value; secondly, the analysis sheds light on the data-related practices and data literacy skills of journalists working in this field; and thirdly, it extends the line of academic work analysing major national or international data journalism competitions (Chaparro-Domínguez & Díaz-Campo, 2021; Ojo & Heravi, 2018; Young et al., 2018). One of the latest valuable pieces in line of this work being a thorough analysis of projects nominated for Data Journalism Awards 2013–2016 (Loosen et al., 2020).

Most of the studies mentioned above use content analysis to look at reoccurring topics for notable data journalism projects, the actors producing them, types of data commonly used etc, proposing ways to define an 'ideal data story' (Ojo & Heravi, 2018) or 'gold standard' (Loosen et al., 2020) for data journalism projects. But they also deal with data verification, transparency and privacy related issues (Chaparro-Domínguez & Díaz-Campo, 2021). The analysis of Sigma Awards 2021 continues this work, offering a recent base of comparison, but also adds to previous studies, because this novel dataset offers a valuable opportunity to map the problems journalists faced while completing their projects – as they described when presenting their finished works.

To sum up, looking at an international dataset of data journalism projects related to COVID-19, offers valuable information for a variety of audiences: media researchers, data holders, journalists and universities. An overview of which COVID-19-related topics were most often covered and on what type of data these projects were based (statistics, medical expertise, etc.) (RQ1) helps comprehend and describe the role journalists took during the pandemic, and the type of data used offers input to further discuss the level of data literacy the journalists currently have, thus providing valuable practical information for educators of journalism updating academic curriculum. This insight is strengthened further by analysing the hardships data journalists encountered while working on their projects (RQ4) – once again reflecting the problems with data, but also reflecting on the skills they needed to work with this data.

Looking at how many and what types of data sources were named in the projects (RQ2) and how many and what types of sources were used in the projects to comment on an aspect of the data presented in the project (RQ3), helps further explore the notion, that data journalism is largely determined by public datasets. The dataset at hand uniquely brings together projects with a united theme. The focus on the shared topic of reporting COVID-19 provides an opportunity for practitioners from different fields – researchers, journalists and data providers – to look at ways data is collected and presented to the public and how to improve the quality of this information.

In addition to focusing on these questions, an overview that maps the Sigma Awards competition of the remarkable pandemic years for future researchers will be given of the background variables of the projects.

METHOD

The dataset for this study is projects submitted to Sigma Awards 2021 and shortlisted by the competition jury (N=140). The projects were submitted as single projects or as portfolios – both could consist of several articles, links to web portals, project-related databases, data visualisations and presentations. For study purposes, sub-sampling was used and only projects related to COVID-19 were selected for further analysis (using keywords: "COVID-19", "corona", "coronavirus", "virus" and/or "pandemic"). The sub-sampling was based on submission info, which is publicly available on the Sigma webpage through the projects database section and is downloadable in .xlsx format via GitHub (*The Sigma Awards Database.*, 2023). Links leading to the projects are also included in the submission info. This process led to the shortlisting of topical pandemic-related projects (n=73). These projects were analysed using manual content analysis, a central method in communication research (Krippendorff, 2018; Lacy et al., 2015).

As mentioned, each individual project or portfolio in the dataset presented one or several weblinks leading to a pandemic-related data journalistic piece. In each case, the initial weblink was used to determine the unit being coded – unless that link led to a presentation about the project, in which case the next link leading to the project was used. Applicants were able to enter multiple projects in the competition, which enabled a few data journalists despite being a member of a project team, to submit their personal portfolios. In these instances, if the story was already coded in the single project category, the first available unique weblink was used to find another pandemic-related piece from that portfolio.

Most of the projects presented to the contest were written in or translated into English. In the case of the project being in a language other than English (for example Chinese or Spanish), the project was read and coded with the help of Google Translate. The length of the projects varied greatly as did their types: from infographics accompanied by a few lines of explanatory text, to feature stories that would take up several pages in print. To conduct the content analysis, the whole text corpus was as an initial step thoroughly read by the author. Next, the author created a preliminary codebook, which was tested on a sample of the dataset for necessary corrections in categories and descriptions. The results were recorded in a tabular format following the codes identified by the codebook. After the pilot coding, the codebook was reviewed and discussed with a second researcher. The author and the second researcher then randomly selected 9 of 73 (approximately 12%) of the shortlisted projects and coded each of them. A comparison of recorded results, tracing differences in codes attributed and then calculating the number of concurrences, showed an intercoder reliability of 88%. The author completed the rest of the coding, after which the dataset was double-checked for possible coding errors.

As 'data' and 'sources' are keys to the codebook and presented analysis, a short clarification of the terms is necessary. Both terms are used throughout the codebook to mark categories – the number of data sources in the project at hand, type(s) of named data sources, type of data used for the project, number of named sources used to comment on an aspect of the data presented as well as the type of these sources. 'Data' is understood here as quantitative information, such as different types of statistics or other numerical knowledge, but because the topic, COVID-19, deals largely with medical expertise, qualitative estimates phrased as "frequent", "large amount", "a little", etc. were also valid for coding. 'Data sources' are people or institutions accredited by name in the project as owners of the data. If the source was referred to by anonymous or unaffiliated 'experts', 'scientists', etc., the source was coded as "not indicated". This was also done in the case of unspecified hyperlinks tied to phrases from the text and acting as additional sources directing outwards from the story.

This analysis also looked at named sources used to comment on an aspect of the data presented in the COVID-19 project. To be coded in this category, a clear connection to the data used for the journalistic piece was sought. This was done by evaluating the way, in which each source was used by the journalist: was it to illustrate the story or give context to specific data?

Journalists who submitted a single project (n=42) were also asked by Sigma to describe the hardships encountered during their work process. These texts, written in their own words, were read to pinpoint and code defined problems. These data collected by Sigma present a valuable opportunity to get "behind the scenes" of data journalism work and systematise and list the main hardships encountered while reporting the pandemic.

Finally, background variables – project title, type of project (single or portfolio), geographical location of the submitter(s) and their organization size (big or small) – were coded to add general context concerning the Sigma Awards competition.

RESULTS

OVERVIEW OF TOPIC AND FOCUS

Overall, the projects in the COVID-19 dataset were mainly (74%) from big newsrooms – where 'big' means consisting of 35 or more journalists (including freelancers and contractors) or done collaboratively. This trend has been noted in international data journalism competitions before – projects found to name over five people as authors or contributors (Loosen et al., 2020). This can be explained by the complex and labour-intensive nature of data journalism projects in general, but even more so in the case of projects chosen to compete for an international award. The dominance of bigger newsrooms is visible, with *Reuters*, *The New York Times* and *The Economist* submitting multiple projects to Sigma Awards. Large media houses have the necessary resources and editorial commitment to invest in cross-disciplinary data teams; however, smaller, regional newspapers can make up for lack of resources via the skilful use of sophisticated tools and approaches (Young et al., 2018).

Geographically speaking, the competition of 2021 indicates that data journalism is gaining momentum worldwide. An analysis of Data Journalism Awards 2013–2016 shows that nearly half of the nominees came from the United States (Loosen et al., 2020). The COVID-19 dataset in the Sigma Awards did have the top three nominee regions: 34% of nominees (from North America), 26% (Europe and 12% (Asia).

Analysing the focus and topic of the projects, two trends emerged. Looking at Table 1 below, the majority of projects focused on the national aspect of the pandemic. This was also supported by the choice of topic: over half of the projects reported on the victims of COVID-19, offering the audience the latest information on infection rate, death toll, etc.

Table 1. Geographical focus and main topic of data journalism projects submitted to Sigma Awards 2021 (%, multiple coding)

Geographical focus (%)		Most common topics (%)	
National aspects of the pandemic	63	COVID-19 victims	52
Global aspects of the pandemic	23	Geographical spread of the pandemic	36
Local aspects of the pandemic	14	Explaining the nature of the virus	15
Semi-global aspects of the pandemic	8	Short-term preventative measures	15
Non-specific to a location	7	Long-term preventative measures	15
		Current state of the medical sector and/or medical care issues	14
		COVID-19 testing	12

Source: Author

DATA SOURCES AND TYPES

Most of the projects named either 2 data sources (22%) or 10+ (26%), which shows the dual nature of the projects. On one hand, COVID-trackers presenting dashboard maps and graphs (these could be built on data from a couple of sources such as the National Ministry of Health and the European Centre for Disease Prevention and Control). On the other hand, long in-depth feature stories (e.g. looking at which US cities have the biggest racial gaps in access to COVID-19 testing or how overcrowded housing spreads the virus among essential and service workers). Consequently, a third of the projects (33%) coded did not involve sources commenting on the data at hand, as so many of the projects were intended as a quick overview of statistics, where the journalistic text was dominated by visualisations (e.g. the coronavirus map of *The New York Times*).

Most of the projects relied on COVID-19-related healthcare statistics and/or medical expertise – one or both of which were used in 66% of projects coded. Thus, the pattern was to cite official institutions (named as data sources in 66% of the projects). In almost half of the cases, the source could be specified as an organisation dealing with health – such as national health ministries, state health departments, etc. Frequently, data also came from national statistics bureaus (e.g., the US Census Bureau, the UK Office for National Statistics, the Brazilian Institute of Geography and Statistics).

The second most popular group involved universities, scientists and academics, think tanks (most often Johns Hopkins University, but also Oxford University,

University of Washington, etc.). Combined with the citing of scientific journals, these groups comprised 17% of all named data sources.

As for the other half of the data sources – the media and journalists emerge. Numerous projects cited their own company as a source, referring to their own dataset, collated data and analysis (e.g., collating data from Google Maps and location of schools to determine how many businesses a new government measure affects). Additionally, some data were quoted from other media – news stories and interviews – as well as from social media posts (e.g., projects covering pandemic disinformation). Together, these two categories comprise 23% of the data sources cited by name. The use of other data sources, such as non-commercial organisations, private companies or clinicians, was rare compared to those mentioned above.

Although the majority of projects relied on healthcare-related statistics, using multiple types of data was common (44%) – a result in line with data journalism research from previous years (Loosen et al., 2020). A typical example of this is healthcare statistics combined with census data to show pandemic numbers in relation to population numbers across regions. This paper does not focus on the visualisations used in the projects, but it is safe to say that dashboard maps collating data on COVID-19 were highly popular (Koch, 2021).

Table 2. Type of data used in data journalism projects submitted to Sigma Awards 2021 (%, multiple coding)

Most frequent type of data, %	
Healthcare-related statistics about COVID-19 situation	60
Medical expertise	16
Census data	15
Results of a poll/questionnaire/survey	12

Source: Author

Against this background, several examples of rarer data are noted. Examples are obituaries, eviction filings, fact checks performed per month or emoji usage data from messaging apps. Often, the data journalistic project concluded with a reference list citing data sources. There was only 1 case out of 73 coded where the journalistic piece did not cite a source.

As for other types of sources, i.e., those used to comment on an aspect of the data presented – the most common were researchers. Scientists were relied upon to offer expert knowledge on a variety of COVID-19-related issues from medical knowhow to information warfare strategy. Thus, they made up roughly 34% of the coded data sources that could be identified through their name. In second place were public institutions or their spokespersons (19%), which were most often used when the focus of the story was related to issues of the state, such as government spending on medical supplies, creating testing access for the public, etc.

HARDSHIPS ENCOUNTERED BY PROJECT TEAMS

All the authors of single projects were given an extra question by the Sigma Awards team. When submitting their work, they were asked about the difficulties encountered during their work process. This provides an opportunity to further look at 42 single projects dealing with COVID-19.

Analysing the difficulties described, most reoccurring problems can be divided into three larger categories as shown below.

Table 3. Most common problem sets in single projects submitted to Sigma Awards 2021

Specific characteristics of project	For example: time consuming, sensitivity, legal issues, security reasons.
Issues with data	• Lack of available data, research • Problems obtaining data from data holders • Inconsistent distribution timewise by data holders of available data • False data, imperfect data, disinformation • Verifying the quality of available data, checking (changes in) data • Language or format in which available data is presented, messy data • Collating gathered data, creating algorithms • Technical challenges keeping project available to public, challenges updating data
Prepping project for public	Technical challenges working with data before making project public Visualising data for projects Large volume of data, size of project

Source: Author

Most often the problems described fell into the first two categories – either specific characteristics of the project (most often about 'time consuming') or issues with data. Either or both of these arose in most single projects.

The time consuming aspect was often associated with other problems, such as obtaining the data (e.g., a project pointing out signs of corruption involving public spending on respirators, a platform publishing data on illicit wildlife trade). Another feature of the time consuming aspect was the sheer volume of data the team had to sort and process: individual microdata of more than 15 million deaths in one project, hundreds of thousands of tweets for another, more than 9000 fact-checking reports for a third project, etc. Of course, an international data journalism competition such as Sigma Awards is prone to receive submissions where the projects are resource-consuming timewise, need more human resources working on them, etc. However, the datafied and complex topic of COVID-19 itself dictated the workload. For example, a project where data journalists worked side-by-side with mathematical modelling academics and epidemiologists for several months with the aim of visualising the effects of vaccinations.

An interesting find was the frequent mention of journalists having trouble obtaining data from data holders and data presented in a 'messy' or unsuitable or uncomfortable language or format to work with. The latter was especially

mentioned by teams working with COVID-19 healthcare data. Data holders, primarily official channels like national health departments, were mentioned to have released some data for a few weeks and then have stopped and changed data definitions or publication times. Relevant data were also frequently scattered across several websites and platforms, and several teams pointed out that healthcare data were released in an unsuitable format – such as .jpg charts – which led to data journalists spending even more time rewriting the data manually themselves.

Data handling issues were mentioned frequently by the authors of the projects – the most popular of them being a) verifying the quality of available data and checking the data for changes; b) collating the gathered data and creating algorithms for the project.

Working on the COVID-19 projects, several teams mentioned the sensitivity of the topic. As described in a recent study looking at UK journalists and how they managed their COVID-19 trauma: "On one level the journalists /--/ had never been more separate from their sources; on another level, they had never been closer" (Jukes et al., 2021, p. 15). Finding ways to visualise death tolls was a journalistic challenge – one of the 2021 winners, the project entitled 'No Epicentro', described the situation: "the numbers were faceless". 'No Epicentro' combined census data with the number of COVID-19 deaths to create an interactive simulation placing all the deaths in the neighbourhood of the reader, thus creating closeness to the tragedy.

A variety of other hardships were described but none of them as frequently as those mentioned above. Examples of this would be financial issues (e.g., lack of funding for the project, having to find funds or unexpected financial expenses), establishing relationships with potential sources, the authors feeling the pressure of time while working on their project and the lack of people working on the project.

DISCUSSION AND CONCLUSIONS

Looking at the various data journalism projects submitted to Sigma Awards 2021, there is no doubt that a great deal of data journalism work has, since 2020, been focused on COVID-19 issues. Analysis shows that 73 of the 140 projects shortlisted for Sigma Awards 2021 had an angle related to COVID-19. It's a topic where data journalism can prove its worth to the audience – health being a heavily datafied area on any given day – and data journalism "with its ambition to use statistics and visualizations for precise and careful investigations" seems "particularly well equipped for covering COVID-19" (Pentzold et al., 2021, pg. 5). The pandemic offered the chance for data journalists to take on the role of knowledge brokers and raise public awareness.

Hence, most of the COVID-19-related projects focused on the national aspect of the pandemic. Although the health crisis is worldwide, the top priority for journalists has been to report accurate and balanced information to their home audiences, thus geographical closeness is the central aspect. Furthermore, the virus has had an immediate or indirect effect on numerous areas of all societies. This is echoed in the works submitted to Sigma Awards, whose topics ranged greatly from government spending to the spread of fake news, from homelessness to change in emoji usage, from virtual memorials to wildlife crime, etc. The variety of genres and story angles support the claim that health news as a genre has always carried elements of "both hard (government decisions about public health strategies) and soft (how to take care of one's personal health) forms of journalism" (Hanusch, 2022, p. 1136).

The analysis shows (RQ1) that the most reoccurring theme was COVID-19's infection rate and death toll – covered by over half of the projects – and this topic was often combined with reporting the geographical spread of the virus. Most often, these specific projects can be described as COVID-19 monitors – quick-to-grasp visualizations of healthcare-related statistics, which in several cases were regularly updated by the data teams, wishing to provide their audience accurate information on the unfolding crisis. Logically, the most heavily used type of data was healthcare-related statistics about COVID-19: drawing on measured values has been a standard combination in data journalism projects covering health issues (Loosen et al., 2020). These findings point at several aspects of data journalistic practices in the newsrooms. First, that journalists reporting COVID-19 relied on data in the form of statistics – and that the role of the project was to visualise the data in an easily readable and informative way. Thus, a data journalist needs not only to have a varied skill-set to make sense of the data, but must also be a visual thinker. This skill profile is something to be considered by those designing curriculum to train future journalists.

As for the hypothesis that the pandemic has further solidified government bodies, universities and other research institutes as data sources, because national healthcare statistics are most often gathered by official institutions, the connection is evident. Analysing the dataset (RQ2), while the most dominant type of cited data source was indeed official institutions like national health ministries or state health departments, in second place were universities, scientists, think tanks and scientific journals. It is evident that Johns Hopkins University has been a great source for data journalists due to their continuously updated COVID-19 dashboard, which collates healthcare data from numerous databases. Cited by name in every fifth project, Johns Hopkins University released their global COVID-19 tracker map – one of the first available – in January 2020 (Dong et al., 2020) and their effort has since grown to gather data from more than 260 sources (*Johns Hopkins University*, 2022). TIME magazine named Johns

Hopkins University the "de facto clearinghouse for pandemic stats", listing it as one of the top inventions of 2020 (Korn, 2020). The data from the Sigma Awards 2021 emphasizes that it is best to serve the public with fast and accurate information in a time of crisis such as COVID-19. To do this, societies would benefit from open discussions between journalists, official bodies and research institutes, focusing on efficient ways of data collection and distribution.

As for the number of named data sources cited, two groups emerged from analysis of the dataset. First, projects citing one to two data sources by name; and second, projects citing more than 10. This in an indicator of the variety of genres covered by data journalists – from short news pieces to in-depth feature stories. In the case of COVID-19, visualizations often proved more important than text, as is evident in a variety of web-based dashboards.

Likely due to the latter, the data in the projects were often presented without commentary on additional sources. As for the projects that added comments concerning the data presented, the most dominant group by far was universities, scientists and academics and think tanks (RQ3) – these were identified by name in roughly a third of the projects. Thus, the dataset underlines the need for public speaking skills for the scientific community.

The dataset also provided an opportunity to look at the difficulties journalists encountered while working on their projects (RQ4). The conclusions presented here are drawn from brief written information provided by the journalists themselves – researching this area further by means of structured interviews would be the next step forward.

Most often the problems described by the data journalism teams fell into two categories – the specific characteristics of the project at hand (most often time consuming) or issues with data. Amplifying the need for a data journalist to use data literacy skills, the problems with data most often had to do with verifying their quality or checking them for changes, but also collating them. Adding to the problem, epidemiologists and other researchers themselves struggled with data issues that, if addressed, could threaten the validity of their results (Stoto et al., 2022). This pressing need to make sense of data falls in line with how Marconi (2020) describes changes facing journalists working in newsrooms: they are also asked to be technologists, making use of new tools at their disposal. Analytical skills to look at data, performing data-mining and generating appealing data-driven projects are necessary traits for current and future data journalists, "new entrants in an increasingly digitised field" (Zhang & Wang, 2022, p. 1128), creating further expectations for journalism education. As explained above, data journalists needs a varied skill set: skills of a data analyst, but also story telling and visual thinking.

Another common problem rising from Sigma Awards dataset, as well as validating the technical aspects of the job, were journalists' mentions of creating

algorithms for the project as well as the technical challenges in keeping the project available to the public and updating the data. At least one of the difficulties listed above was mentioned in every other submission text for COVID-19-related single projects.

An interesting find was the frequent mention of journalists having trouble obtaining data from data holders and data presented in a "messy" format. Again, this highlights the complex relationship between the media and data sources in the time of datafication, which should be a vital topic for future media research. In the cases of official institutions where public interest is of the highest priority, an open dialogue with journalists concerning data formatting and availability seems especially beneficial for all parties.

Finally, some limitations of the current study should be pointed out. The entries of the Sigma Awards, especially the shortlisted projects, present the highest level of data journalism currently done in the world. Most of the projects analysed came from big newsrooms and large teams – this could impact the complexity of the projects (number of sources used, intricacy of theme etc.), but also the problems they encountered, because of a greater potential variety of data-related skills in a bigger team.

Beyond doubt, coverage of COVID-19 has amplified the public need for fast and accurate data journalism. But how this stressful time has affected journalists themselves is a topic worth researching further – worldwide and on a skills assessment level.

REFERENCES

Beiler, M., Irmer, F., & Breda, A. (2020). Data Journalism at German Newspapers and Public Broadcasters: A Quantitative Survey of Structures, Contents and Perceptions. *Journalism Studies*, *21*, 1–19. https://doi.org/10.1080/1461670X.2020.1772855

Borges-Rey, E. (2020). Towards an epistemology of data journalism in the devolved nations of the United Kingdom: Changes and continuities in materiality, performativity and reflexivity. *Journalism*, *21*(7), 915–932. https://doi.org/10.1177/1464884917693864

Casero-Ripolles, A. (2020). Impact of Covid-19 on the media system. Communicative and democratic consequences of news consumption during the outbreak. *El Profesional de La Información*, *29*(2). https://doi.org/10.3145/epi.2020.mar.23

Catalan-Matamoros, D., & Elías, C. (2020). Vaccine Hesitancy in the Age of Coronavirus and Fake News: Analysis of Journalistic Sources in the Spanish Quality Press. *International Journal of Environmental Research and Public Health*, *17*(21), 8136. https://doi.org/10.3390/ijerph17218136

Chaparro-Domínguez, M.-Á., & Díaz-Campo, J. (2021). Data Journalism and Ethics: Best Practices in the Winning Projects (DJA, OJA and Sigma Awards). *Journalism Practice*. Scopus. https://doi.org/10.1080/17512786.2021.1981773

Cushion, S., Lewis, J., Sambrook, R., & Callaghan, R. (2016). *Impartiality Review: BBC Reporting of Statistics Report on qualitative research with the BBC audience*. http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/our_work/stats_impartiality/content_analysis.pdf

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Engebretsen, M., Kennedy, H., & Weber, W. (2018). Data visualisation in Scandinavian newsrooms: Emerging trends in journalistic visualisation practices. *Nordicom Review*.

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*, *5*(2), 205395171878631. https://doi.org/10.1177/2053951718786316

Hanusch, F. (2022). Change and Continuity in Digital Journalism: The Covid-19 Pandemic as Situational Context for Broader Arguments about the Field. *Digital Journalism*, *10*(6), 1135–1140. https://doi.org/10.1080/21670811.2022.2092020

Johns Hopkins University. (2022). Johns Hopkins Coronavirus Resource Center. https://coronavirus.jhu.edu/about

Jukes, S., Fowler-Watt, K., & Rees, G. (2021). Reporting the Covid-19 Pandemic: Trauma on Our Own Doorstep. *Digital Journalism*, *0*(0), 1–18. https://doi.org/10.1080/21670811.2021.1965489

Koch, T. (2021). Welcome to the revolution: COVID-19 and the democratization of spatial-temporal data. *Patterns*, *2*(7), 100272. https://doi.org/10.1016/j.patter.2021.100272

Korn, M. (2020). *2020's Go-To Data Source: Johns Hopkins Coronavirus Resource Center*. Time. https://time.com/collection/best-inventions-2020/5911434/johns-hopkins-coronavirus-resource-center/

Krippendorff, K. (2018). *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

Lacy, S., Watson, B. R., Riffe, D., & Lovejoy, J. (2015). Issues and Best Practices in Content Analysis. *Journalism & Mass Communication Quarterly*, *92*(4), 791–811. https://doi.org/10.1177/1077699015607338

Lewis, S. C. (2020). The Objects and Objectives of Journalism Research During the Coronavirus Pandemic and Beyond. *Digital Journalism*, *8*(5), 681–689. https://doi.org/10.1080/21670811.2020.1773292

Loosen, W., Reimer, J., & De Silva-Schmidt, F. (2020). Data-driven reporting: An on-going (r)evolution? An analysis of projects nominated for the Data Journalism Awards 2013–2016. *Journalism*, *21*(9), 1246–1263. Scopus. https://doi.org/10.1177/1464884917735691

MacDonald, N. E. (2021). COVID-19, public health and constructive journalism in Canada. *Canadian Journal of Public Health*, *112*(2), 179–182. https://doi.org/10.17269/s41997-021-00494-8

Makri, A. (2021). What do journalists say about covering science during the COVID-19 pandemic? *Nature Medicine*, *27*(1), 17–20. https://doi.org/10.1038/s41591-020-01207-3

Marconi, F. (2020). *Newsmakers: Artificial Intelligence and the Future of Journalism*. Columbia University Press.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work, and Think*. Houghton Mifflin Harcourt.

Nisbet, M. C., & Fahy, D. (2015). The Need for Knowledge-Based Journalism in Politicized Science Debates. *The ANNALS of the American Academy of Political and Social Science*, *658*(1), 223–234. https://doi.org/10.1177/0002716214559887

OECD. (2016). *Skills Matter: Further Results from the Survey of Adult Skills*. OECD. https://doi.org/10.1787/9789264258051-en

Ojo, A., & Heravi, B. (2018). Patterns in Award Winning Data Storytelling. *Digital Journalism*, *6*(6), 693–718. https://doi.org/10.1080/21670811.2017.1403291

Parasie, S., & Dagiral, E. (2013). Data-driven journalism and the public good: "Computer-assisted-reporters" and "programmer-journalists" in Chicago. *New Media & Society*, *15*(6), 853–871. https://doi.org/10.1177/1461444812463345

Pentzold, C., Fechner, D. J., & Zuber, C. (2021). "Flatten the Curve": Data-Driven Projections and the Journalistic Brokering of Knowledge during the COVID-19 Crisis. *Digital Journalism*, *0*(0), 1–24. https://doi.org/10.1080/21670811.2021.1950018

Porlezza, C., & Splendore, S. (2019). From Open Journalism to Closed Data: Data Journalism in Italy. *Digital Journalism*, *7*, 1–23. https://doi.org/10.1080/21670811.2019.1657778

Quandt, T., & Wahl-Jorgensen, K. (2022). The Coronavirus Pandemic and the Transformation of (Digital) Journalism. *Digital Journalism*, *10*(6), 923–929. https://doi.org/10.1080/21670811.2022.2090018

Royal, C. (2012). The Journalist as Programmer: A Case Study of The New York Times Interactive News Technology Department. *ISOJ Journal*, *Volume 2, Number 1, Spring*, 5–24.

Stalph, F. (2018). Classifying Data Journalism. A Content Analysis of Data-Driven Stories. *Journalism Practice*, *Vol. 12, No. 10*, 1332–1350. https://doi.org/10.1080/17512786.2017.1386583

Stephan, M., Register, J., Reinke, L., Robinson, C., Pugalenthi, P., & Pugalee, D. (2021). People use math as a weapon: Critical mathematics consciousness in the time of COVID-19. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-021-10062-z

Stoto, M. A., Woolverton, A., Kraemer, J., Barlow, P., & Clarke, M. (2022). COVID-19 data are messy: Analytic methods for rigorous impact analyses with imperfect data. *Globalization and Health*, *18*(1), 2. https://doi.org/10.1186/s12992-021-00795-0

The Sigma Awards database. (2023). GitHub. https://github.com/Sigma-Awards

Van Witsen, A. (2020). How Daily Journalists Use Numbers and Statistics: The Case of Global Average Temperature. *Journalism Practice*, *14*(9), 1047–1065. Scopus. https://doi.org/10.1080/17512786.2019.1682944

Widholm, A., & Appelgren, E. (2020). A softer kind of hard news? Data journalism and the digital renewal of public service news in Sweden. *New Media & Society*, 1461444820975411. https://doi.org/10.1177/1461444820975411

Wihbey, J. (2017). Journalists' Use of Knowledge in an Online World: Examining reporting habits, sourcing practices and institutional norms. *Journalism Practice*, *11*(10), 1267–1282. Scopus. https://doi.org/10.1080/17512786.2016.1249004

Young, M. L., Hermida, A., & Fulda, J. (2018). What Makes for Great Data Journalism? *Journalism Practice*, *12*(1), 115–135. https://doi.org/10.1080/17512786.2016.1270171

Zhang, S., & Wang, Q. (2022). Refracting the Pandemic: A Field Theory Approach to Chinese Journalists' Sourcing Options in the Age of COVID-19. *Digital Journalism*, *10*(6), 1115–1134. https://doi.org/10.1080/21670811.2022.2029521