MEDIATIZATION STUDIES 2/2018 DOI: 10.17951/ms.2018.2.103-123

FLORIAN MEISSNER

Department of Communication Sciences, Institute for Social Sciences, Heinrich Heine University Düsseldorf

florian.meissner@hhu.de

GERRET VON NORDHEIM

Institute for Journalism, TU Dortmund University

gerret.vonnordheim@tu-dortmund.de

Exploration of a fragmented discourse. Privacy and data security in *Süddeutsche Zeitung*: 2007–2017

Abstract: The goal of this exploratory case study is to identify different facets of news reporting on surveillance, privacy and data security, and more specifically, how risks in this context are portrayed. The theoretical foundation consists of two elements: 1) the concept of mediatized risk culture, and 2) the discursive arena model of risk communication, which provides the normative background for assessing news reporting. A text-mining approach (topic modeling) is applied to analyze relevant coverage of the German quality newspaper Süddeutsche Zeitung. The study yields a total of seven topics which belong to three categories: violation of privacy norms, power and law enforcement, and datafication. The results show that despite the de-mystification of digital technology after the Snowden leaks, coverage has recently become more affirmative and less focused on risk. We conclude that this may indicate a normalization of mass surveillance and data harvesting even in Germany, a society which traditionally values privacy. In order to add more context to our findings, however, further qualitative analyses were needed. The paper serves as a starting point for further research on media reporting of surveillance, privacy and data security.

Keywords: Internet, digitization, privacy, data security, information security, surveillance, Snowden, risk communication, journalism.

Introduction

With the recent revelations concerning the social network giant Facebook and the data analytics firm Cambridge Analytica, the debate on privacy and data security has gained new momentum. The improper exploitation of at least 87 million *Facebook* user accounts by *Cambridge Analytica* is just one prominent example that the concept of privacy is at stake in an increasingly data-driven world. While politicians are interested in big data for both campaigning and security policy, private organizations use and sell data for commercial purposes. What does this mean for our privacy and the security of our personal data? These questions have been subject to public scrutiny not only since the *Cambridge Analytica* scandal. The revelations of former CIA contractor Edward Snowden in 2013 concerning the bulk surveillance by the US intelligence agency NSA in particular form another landmark event in this debate.

In this paper, we apply a text-mining approach (topic modeling) to obtain a comprehensive picture of the discursive position of surveillance and (counter-)discourses of digital privacy and data protection in the Süddeutsche Zeitung[1], Germany's leading quality newspaper, between 2007 and 2017. The center-left daily is one of the most influential agenda-setting media outlets in Germany. It is also known for its investigative stories, including reporting on surveillance and data scandals. There are several international studies that have used the Süddeutsche Zeitung as the exemplar of the German press (Nordheim, Boczek, Koppers 2018; Schäfer, Ivanova, Schmidt 2011). The paper therefore serves as an appropriate case study for one of the first exploratory content analyses concerning the public debate about privacy and data security before and after the Snowden revelations. Another reason is that as a center-left newspaper that often stresses civil rights, one could expect to find substantial awareness of risk to privacy and data security.

The study touches upon the often ambiguous character of emerging, big-scale technologies (Beck 1992; 2009), which is also true for digitization: While its enormous advantages are indisputable, there are also considerable risks inherent in this ongoing technological revolution, such as the increasing vulnerability of privacy and data security. These risks have been debated since the early days of the World Wide Web. For instance, studies conducted in Germany by Beck and Vowe (1995) as well as Rössler (2001) showed that newspaper reporting of the technological advancements and their social and political consequences was dominated by euphoric assessments, but also included skeptical or even apocalyptical ones. Studies from the US and Canada yielded similar findings (Patnode 2003; Sklar 1997). As we know from later research, critical views on digitization gained more prominence in the early 2000s, including growing concerns about privacy and security issues (Oggolder 2015; Zeller,

[1] Average daily circulation in the 4th quarter of 2017: 367,235. See http://www.ivw.eu/aw/print/qa/titel/1221, last accessed on April 10, 2018.

Wolling, Porten-Cheé 2010). However, it was not until 2013 when the whistleblower Edward Snowden revealed the bulk surveillance programs of the US intelligence agency, NSA, that the internet was de-mystified (Möller 2017). At that time, privacy and data security became a major societal and academic concern. Due to the strong focus on the communicative negotiation of privacy vs. security (see next chapter), however, previous research offers an incomplete picture of this complex thematic area. Studies that connect the various facets of privacy and data retention are still missing. For instance, little is known about how they are discussed from an economic or user-oriented perspective, and through which lenses involved actors (administrations, intelligence agencies, private companies…) are viewed.

The research questions of this study are: 1) which facets do news reporting of surveillance, privacy and data security comprise in the *Süddeutsche Zeitung*? 2) Specifically, how are risks to privacy and data security portrayed? The results of this exploratory case study are expected to lay the groundwork for a bigger and international sample.

For the purpose of this analysis, we tentatively conceptualize privacy as a state where "people have control over the disclosure of their personal information" (Budak, Rajh, Recher, 2017, p. 37). Data security (or information security) is understood as the protection of data in terms of confidentiality, accessibility, and integrity (BSI 2018), while surveillance describes technological means to obtain personal and meta-data from the digital traces of people (Čas et al. 2017).

The privacy versus security "trade-off"

Governments and authorities across the globe find themselves in an ambiguous role as the internet is both an object and tool of security policy (Deibert 2016). It is an *object* of security policy because it is a critical infrastructure that modern societies increasingly rely and depend on. It is a *tool* of security policy because surveillance is widely considered the key to fight organized crime and terrorism. The Snowden revelations sparked an international debate about the tension field surrounding surveillance, privacy, and security (Čas et al. 2017). The trend towards data-driven security policy, however, seems to be unbroken. In this context, political and media actors have commonly framed security and privacy as a "trade-off" (ibid.). Scholars like Čas et al. (2017, p. 4–5) criticize this as a narrowed view where digital surveillance is presented as the prime solution, whereas privacy is often framed as an obstacle to security instead of conceptualizing it as an element of (individual) security.

Triggered by the Snowden leaks, the negotiation of security versus privacy has also become a concern for a variety of communication scholars. For instance, Johnson (2016) investigated the responses of leading US media organizations to the revelations and found that opinion pieces comprised diverse views about ethics, legitimacy and necessity of government surveillance. Wäscher (2016) looked at the communicative

strategies of anti-surveillance movements based in the United States and how these strategies opened discursive spaces in public debate. Despite such activities, however, Dencik und Cable (2017) argue that citizens widely show resignation, not consent, when it comes to the lack of knowledge and control over what happens to personal data in an increasingly digitized world. The authors have dubbed this phenomenon "surveillance realism". Thorsen (2016) investigated post-Snowden news coverage by *The Guardian* and *The New York Times* concerning encryption of private online communication. He found that the reports provided tech giants such as Google, Facebook, Apple, Microsoft and others with a platform to exonerate themselves from privacy-related allegations rather than offering helpful information for ordinary users.

Most research in this area, however, is closely linked to the debate in the United States. The next section will therefore look at the specifics of the German debate and relevant historical backgrounds.

The debate about privacy and data security in Germany

It has often been noted that Germans can be notoriously sensitive when it comes to data privacy. Jarvis (2011) has called this sensitivity the "The German Privacy Paradox." Indeed, studies have shown that privacy awareness is higher in Germany as compared to other western countries. For instance, Löblich and Karppinen (2014) analyzed newspaper reporting of internet governance in Germany, Sweden, Finland, and the US. They found that in Germany, the political debate about privacy was most pronounced, and there was also a stronger emphasis on regulation. On the contrary, US newspapers demonstrated a deregulatory attitude and a stronger emphasis on security. Longitudinal studies conducted by Oggolder (2012; 2015) yield further support for the notion that German media pays special attention to privacy–related issues in the digital world.

For a better understanding, one has to look into the historical backgrounds of the debate in Germany. First and foremost, it has to be taken into account that Germany experienced two dictatorships, both of which used intelligence services to spy on their own citizens and violate people's most basic rights. These experiences have shown lasting effects. For instance, a detailed census planned by the West German federal government in 1983 elicited large-scale protests. The Constitutional Court made a groundbreaking judgement which involved a "fundamental right to informational self-determination" (Albers 2017, translated). The census, which was actually conducted four years later, had to be adjusted to the new privacy rules.

Later debates in the reunified Germany included surveillance activities by the domestic intelligence, the Constitutional Protection Agency, and the Data Retention Directive, which stipulates that telecommunication firms must store their customers' personal and communication-related data for a certain period of time to facilitate criminal prosecution (Lüber 2014).

More recently, the focus has shifted to private companies collecting massive amounts of personal data for commercial purposes. The German government has been a driving force behind the current strict data protection policies of the European Union, which are now being implemented in the member states. Paradoxically, when it comes to state surveillance, even the Snowden leaks seem to have had little effect on policies. Dimmroth, Steiger und Schünemann (2017) have outlined that despite public outrage over the revelations, the intelligence cooperation between the US and Germany has been marked by continuity rather than critical reassessment.

Theoretical framework

The theoretical foundation of the study consists of two major elements that we consider relevant to the mediated debate about privacy and data security: 1) the concept of mediatized risk culture, and 2) the discursive model of risk communication.

The concept of mediatized risk culture (Roslyng, Eskjær 2017) is an approach used to capture the way media logic permeates debates about risk. The authors show "how risk is presented as manageable and/or contested and how this, to a certain degree, is a result of a mediatized logic" (ibid, p. 116). As media tends to emphasize conflicts surrounding risk and a lack of controllability, it significantly differs from authorities and scientific organizations. This also influences "the way risk actors engage in the public debate on different media platforms and, possibly, how media institutions may contain a bias in favor of particular risk actors" (ibid, p. 115). The authors furthermore argue that risk representations in the media are influenced by the degree risks resonate with "deep-seated cultural images and narratives" (ibid, p. 126).

The observation that risk perceptions are highly culture-dependent has been made by several scholars before. Beck (2009) argues that the dominant values and norms within a society influence the way how we perceive and deal with risk. In the German context, we can expect that the mentioned historical experiences (see the prior section) have shaped the values and norms that determine how risks to privacy and data security are debated. It is furthermore important to note that risk perceptions are not static, but are subjected to (mediated) processes of negotiation (Keller 2003). This is why media reporting of risk is often considered to have far-reaching political implications.

The discursive arena model of risk communication (Bonfadelli 2004) is a normative approach which attempts to place the negotiation of risk in a democratic context. It is based on the observation that public communication about the risks induced by science and technology often focuses too much on the views of experts. The model thus postulates to include the broad public into the negotiation of risk, especially when large parts of society are (potentially) affected. Thanks to the participatory approach of the discursive arena model, the public debate can be understood as a forum for

different perspectives on risks related to privacy and data security, encourage public scrutiny, and serve as an instrument of early warning, e.g. when state actors or private organizations pursue questionable policies. By involving civil society in the risk communication process, the discursive arena model also aims at sensitizing citizens to risks and addresses their self-responsibility. This is a key prerequisite for privacy and data security in the cyber world.

The discursive arena model, however, is not a naïve approach. It acknowledges that journalistic news reporting of risk often fails to live up to normative expectations. According to Bonfadelli (2004, pp. 296–298), significant shortcomings in this context are a high dependence on the political agenda, limited source plurality, lack of in-house expertise, as well as event-driven reporting (see also Henn, Vowe 2015; Pantti, Wahl-Jorgensen, Cottle 2012). The study presented in this paper allows us to assess whether or not these (or other) problems can be observed in the context of privacy and data security.

Method and sample

The first step of any content analysis is finding a sample that helps to answer the research questions, in this case: Which facets of surveillance, privacy and data security are mirrored in the newspaper articles (RQ1) and, specifically, how are risks to privacy and data security portrayed (RQ2)? Since our study is exploratory and meant to comprise the entire discourse on data security and privacy in the *Süddeutsche Zeitung*, we chose a broad selection criterion. The trade-off of imprecision when using very open search terms is compensated by subsequent topic clustering processes (see below). The search term[2] was applied to a ten-year news period (beginning of 2007 until the end of 2017). As further pre-processing steps, we deleted numbers and punctuation, removed stop words using a list of 612 stop words (see supplement), and transformed all words to lowercase. For our text-mining approach, we used only words that appear more than five times in the corpus. In a second step, we applied the topic clustering algorithm LDA to the subcorpus that was thus generated.

[2] "ueberwachung" (surveillance) AND "NSA" – OR – "ueberwachung" (surveillance) AND "geheimdienste" (intelligence services) – OR – "ueberwachung" (surveillance) AND "Internet" – OR – "ueberwachung" (surveillance) AND "daten" (Data) – OR – "ueberwachung" (surveillance) AND "online" – OR – "privat" AND "daten" (Data) – OR – "privat" (private) AND "internet" (Data) – OR – "privat" (private) AND "online" (data) – OR – "personenbezogene" (private) AND "daten" (Data) – OR – "ausspaehen" (spy) AND "daten" (data) – OR – "persoenliche" (personal) AND "daten" (data) – OR – "datenschutz" (data protection) – OR – "datensicherheit" (data safety). Uppercase and lowercase letters were ignored, and the search terms are pattern-defined so that compounds containing the terms are also included.

The statistical model Latent Dirichlet Allocation (LDA) (Blei, Nag, Jordan 2003) is currently the most popular clustering algorithm for processing large text corpora. It reliably predicts the thematic structure of large (journalistic) text corpora (Jacobi, van Atteveldt, Welbers 2016; Puschmann, Scheffler 2016). LDA offers two main advantages: 1) it allows the researcher to discover patterns and structures across *vast amounts of data* that cannot be processed manually, for example the full coverage of all major online news outlets in a given country over a long period of time, and 2) it is an *unsupervised method*, i.e. the method works independently from the researcher's prior knowledge and thus also detects patterns that were unanticipated (DiMaggio, Nag, Blei 2013, p. 593). This means that theoretical assumptions and hypotheses no longer limit the scope of analysis. However, LDA models certainly have some limitations that must be acknowledged. The bag-of-words assumption is frequently mentioned as a weakness due to its elimination of word order, "specific local context information on semantic relations between words is lost, which otherwise might help interpret deeper meanings and solve ambiguities" (Maier et al. 2018 p. 4). Another limitation inherent in the model is the fixed number of topics, which are not estimated by the model and instead must be determined by the researcher in advance. In addition, LDA is not able to model the evolution of topics. The assumption of a fixed number of topics may be detrimental, for example, for long-term analyses. Finally, one of the most discussed problems is the validation of the LDA model. Different approaches have been suggested, both quantitative and qualitative in nature, and which are beyond the scope of this paper (detailed descriptions can be found in Maier et al. 2018, p. 7–8). In this study, we use three approaches. We investigated: 1) intratopic semantic validity (with intruder tests, heuristic analysis of top words and top articles, and labeling of topics), 2) intertopic validity (analysis of the statistical proximity of the topics by means of clustering), and 3) external validation (analysis of temporal patterns and comparison with external events; see, for example, NSA peak in the results).

We applied LDA using the tosca R-package (Koppers et al. 2018). The LDA produces lists of words called topics. These lists assign a probability value to each word (type) in the corpus (topics are distributions over words). Ordered by these probabilities, they yield lists that, in the best case, activate thematically consistent cognitive patterns to human readers. For instance, a human coder can easily thematically correlate the words *nsa, snowden, usa, bnd*[3]*, obama, intelligence agencies* – it is about the *Snowden* leaks surrounding the *intelligence agency NSA*, the role of the *BND* and of individual political actors such as the then US President, Barack *Obama*. These word lists are generated based on the insight that there is a correlation between similar meanings of linguistic units (words, phrases) and the similarity of their distribution across certain linguistic contexts, gained from empirical corpora (distributional hypothesis, first

[3] The Federal Intelligence Service (German: "Bundesnachrichtendienst", abbreviated: "BND") is the foreign intelligence agency of Germany.

made by Zellig 1954). This characteristic of language relates to the cognition-psychological model of cognitive patterns that enable and organize human thought. The LDA topic clustering process harnesses this characteristic of language by putting words into a thematic context depending on how frequently they occur together in a document. This also connects LDA to concepts from communication science that use cognition-psychological concepts such as the framing approach. According to Entman (1993), a frame comprises up to four elements: problem definition, cause diagnosis, moral judgments, and suggestion of remedies. Several studies have provided evidence that reporting shapes the audience's cognitive frames with regard to science and technology, at least to a certain degree (Cobb 2005; Nisbet, Hart, Myers, Ellithorpe 2013; Wolling, Arlt 2015). Even though frames describe specific patterns that do not necessarily have to reoccur one-to-one in LDA topic patterns, there is a conceptual overlap that enables us to refer to the framing approach in our LDA analysis. Jacobi et al. (2016, p. 92) assert that "if framing devices correspond to specific (latent) patterns of vocabulary use, LDA can capture these classes in specific topics, and as such LDA results can also include the frames used in a corpus of texts" (see also DiMaggio et al. 2013).

In LDA, each word in each text is assigned to a topic (token) – depending on the context, the German word "Bank", in its meaning of 'bench', can be assigned to another topic in an article on soccer than in a report on finances, where "Bank" would mean the same as the English term "bank" as a financial institution. The example from the present analysis shows how topic clustering works:

Topic 1: nsa snowden bnd obama washington edward american intelligence agencies american usa

Topic 7: data protection spd european fdp verdict future green committee brussels data retention Washington Following outrage about mass spying of telephone data, the US Congress initiated a reform of intelligence agency NSA. On Wednesday, the House of Representative justice committee voted for a legal amendment that will prohibit US intelligence agencies from collecting millions of telephone connection data sets from US citizens (…)

Two topics were identified in this document which can be seen as a *distribution over topics*: the NSA/Snowden topic we mentioned, and a topic that apparently revolves around privacy legislation. The top words (words with the highest probability for this topic, here translated into English) include, for example, various German political parties as well as names of controversial laws ("mass data retention") and legal terminology ("verdict"). According to these patterns, the words in the sample article were assigned to the topics. Words like "outrage", "mass spying" and "telephone data" were assigned to the first topic, and terms like "reform" and "legal amendment" to the second one.

Three parameters must be defined for the LDA model (K, α, and β). K must be selected for the number of topics, α, and β control whether documents (as distri-

bution of topics) or topics (as distribution of words) are strongly dominated by individual topics or words, or whether the probability for each element is similar. In simple words, these variables can be used to determine whether the probability distribution tends to be sharp or even.

Since there is no satisfactory measure of the coherence of topics, most studies use qualitative approaches to calibrate the variables and validate the topics (Maier et al. 2018, p. 24). Manual labeling or a review of top words (most representative terms) and top articles (most representative articles, i.e. articles with the largest percentage of words from a certain topic). We supplement this qualitative investigation with the intruder words approach for systematic manual evaluation, suggested by Chang et al. (2009). For a tested topic, they used the list of the top N terms and inserted a random intruder term with high probability from another topic. If coders choose the correct answer (i.e. identify the intruder), it can be assumed that the topic is consistent (see results section).

After a manual review of various variable constellations, we used a LDA with a total number of 11 clusters (K=11), i.e. 11 topics (alpha and beta were each defined as 1/K). From this point on, clustering was automatic and thus with a lesser bias than other content-analytical methods. For the manual interpretation, we subsequently used the top words, top articles and topic evolution over time (assigned words on certain topics relative to the subcorpus). For this procedural dimension, the issue-attention cycles of news reporting as described by Waldherr (2012, 2014) serve as an important point of reference because they highlight the dynamic and often short-term oriented character of news reporting. The approach is also relevant with regard to priming and agenda-setting effects among the audience, which is another important precondition for sensitization.

As a third step, comparing probability distributions across word lists also enabled us to cluster topics at a higher level. Tosca offers a function to '*clusterTopics*', which enables grouping with the help of the Hellinger distance. Here, the probability values of individual words in different topics are compared with each other so that topics similar to one another can be grouped (see Fig. 2).

Findings

Reporting frequency

Figure 1 shows the size of the subcorpus that was formed using the search words. A total of 4,998 texts were identified in the period from the beginning of 2007 until the end of 2017. In the summer of 2013, Edward Snowden's revelations marked a clear peak. Post-Snowden, the frequency of articles on the subjects of data security, privacy etc. remained at a consistently high level compared to the time prior to the NSA leaks.

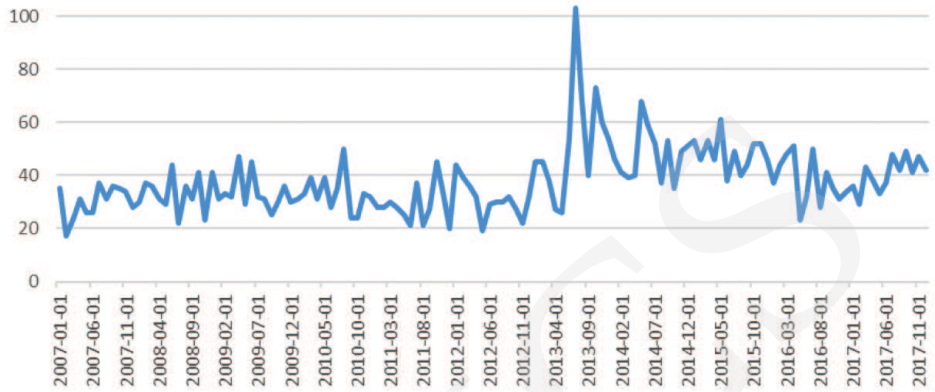

Figure 1. Count of articles in subcorpus over time

This justified the characterization of the Snowden leaks as a "key event" (Rauchenzauner 2008) for the subsequent reporting on surveillance, privacy and data security.

Topic Clustering

Of the 11 generated topics, 10 could be clearly labeled – activating consistent patterns in the coders. The intruder word validation yielded the following result:

Table 1. The *intruderTopics* function of tosca displays a list of five words for each topic. For each test, the user must evaluate eleven lists (with K=11). One or none of the terms in every list is an intruder that must be identified by entering its number (or "0" for no intruder). The representative topic words (top words) and the intruder terms (which are top words from other topics of the same model) are selected randomly and change with each repetition of the test. The test was repeated ten times by two coders (five times each). The table displays the numbers of correctly evaluated lists (i.e., a correctly identified intruder or "0" for no intruder) for each topic.

Topic	1	2	3	4	5	6	7	8	9	10	11
Correct evaluation	9/10	10/10	10/10	9/10	9/10	9/10	7/10	8/10	8/10	10/10	9/10

The quality of LDA resides in its ability to form topics from filler words that are frequently used and thus quite vague, resulting in higher consistency of other topics. In the present LDA, we were able to identify a filler word topic (topic 2). Another characteristic of this method is that it identifies irrelevant texts in relation to the research question – i.e. one can accept low search word precision in favor of comprehensive recall. The imprecision (or, positively worded: openness) of the search word becomes evident, for instance, when a subcorpus features texts about the corporate

data of a *private* bank – the LDA clusters these thematically unrelated words into false positives, which allows us to exclude them later on.

This leaves us with a total of seven topics, which were clustered into three groups using the Hellinger distance (see Figure 2). We consider these three groups as categories of topics that are closely connected to each other. A content analysis of the most representative words and articles of each topic reveals distinct frames that will be discussed subsequently.

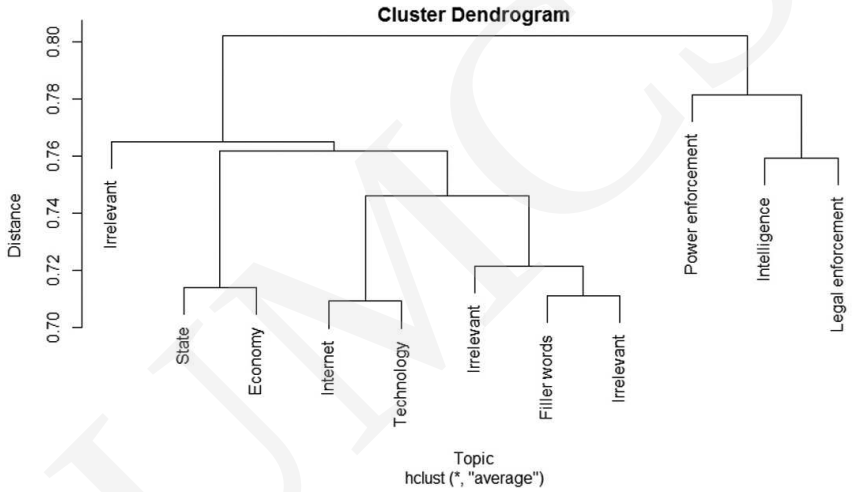

Figure 2. The seven labeled topics are grouped into three broad categories, which are described below under the following headings: Violation of privacy norms (consisting of the topics *economy* and *state*), Power and legal enforcement (*power enforcement*, *intelligence* and *legal enforcement*) and Datafication (*internet* and *technology*).

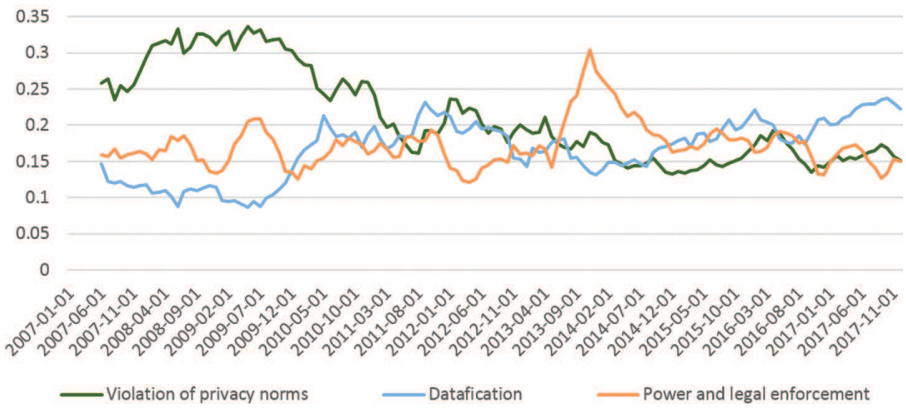

Figure 3. Comparison of topic clusters (sum of single topics, see below), relative to the generated subcorpus (moving average of 6 months).

Violation of privacy norms

The frame category *violation of privacy norms* is about negotiating the limits of economic and governmental practices set by personality rights, in particular, the right to privacy. Associated frames can be generalized as follows: The problem is that by their actions, governmental or economic institutions challenge the rights of citizens (or customers, consumers, or employees). The reasons therefore reside in the specific interests of the acting system: The German Office for the Protection of the Constitution wants to monitor computers (SZ, 24 April 2007: "FDP in North Rhine-Westphalia affronts its own ministers"), corporations are interested in personal data because they want to assess their clients' credit scores (SZ 24 January 2008: "Tricky data"). The frame is usually set up from the perspective of the 'common man' who is the moral victim, looking to privacy or consumer protection legislation for a solution. The state thus plays a double role in this: as the executive branch that wants to enforce mass data collection for reasons of homeland security, and as the judicial branch that is supposed to defend the citizens' rights.

Table 2. The 20 most representative words in each of the two associated topics state and economy (translated).

State	Economy
spd	customers
Privacy	euro
data	data
law	corporations
European	telekom
eu	employee
citizen	railway
fdp	worker
union	banks
european	bank
verdict	privacy
agreement	employees
green	customer
data retention	money
schaeuble	consumers
federal government	employees
committee	information
cdu	cost
Bundestag	insurer
brussels	consumer advocates

The two associated topics clearly show that the attention for economically or governmentally motivated threats to civic rights has decreased relative to the subcorpus

(*figure 1*). A series of spying incidents in 2008, in particular, caused an early peak in the topic 'economy', e.g. SZ, 1 August 2008:

> "Big Brother in the Office
>
> Suspicions are running high in German companies. Each month, new cases of secret surveillance come to light. After Lidl and Telekom, insurance company Gerling now admits to having secretly spied on its employees…"

In the topic *state* the most significant debate is about an amendment to the law that governs the authority of the German Federal Criminal Office in 2008 and the accompanying debate on the retention of mass data. Yet it is apparent that both topics have played an ever-lesser role in the overall discourse since.

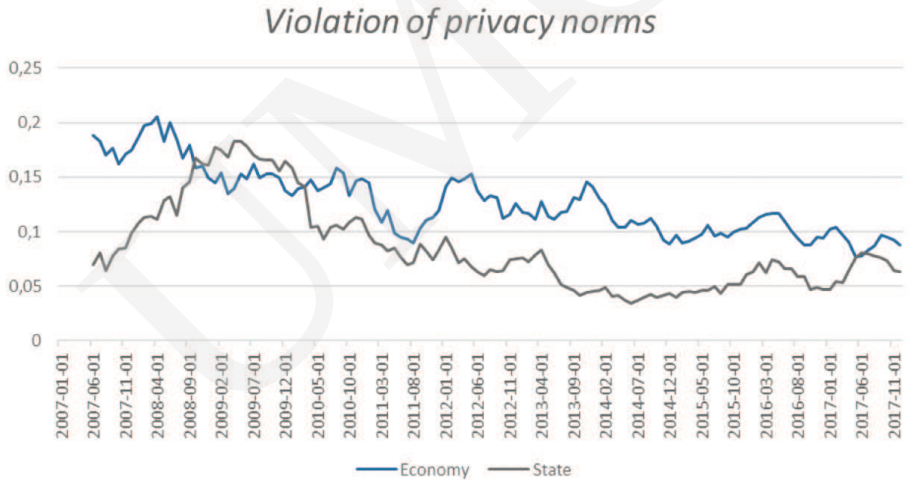

Figure 4. Share of words that were assigned to the topics of *economy* and *state* in the LDA process, relative to the generated subcorpus (moving average of 6 months).

Power and legal enforcement

The clustered topics in this frame category are associated with cognitive patterns about *power and legal enforcement*. In this context, data is: 1) A means of governmental transgression, for instance, in the context of the NSA surveillance apparatus or used against regime critics in countries like Turkey or China; 2) as a means to enforce executive power, for example, to convict pedophile criminals or tax evaders (see the top word "panama" for "panama papers"); or 3) as a weapon in the war on terror. All three functions cannot be cleanly separated – after all, it depends on the individual author's view whether they consider PRISM as a legitimate tool of executive power or an illegitimate abuse thereof. Hence, the discourse negotiates the moral legitimacy

of power gained through data. In this sense, data are tools to punish those who have used unlawful violence (for example, in court cases). At the same time, data itself is increasingly a means to exert violence in disputes. In this frame, the framing of the associated top article is ambivalent as they refer to both problems and remedies.

Table 3. The 20 most representative words in each of the associated topics *intelligence, legal enforcement* and *power enforcement* (translated).

Intelligence	Legal enforcement	Power enforcement
nsa	prosecutor	police
snowden	been	government
usa	investigations	turkey
bnd	court	china
obama	because	country
intelligence agencies	case	policemen
edward	bank	refugees
washington	euro	syria
government	lawyer	authority
merkel	journalists	iran
surveillance	bavaria	chinas
american	panama	perpetrator
american	investigator	is
american	files	paris
intelligence agency	known	terrorists
federal government	man	iraq
chancellor	newspaper	war
wikileaks	papers	officials
barack	edathy	russia
leaks	swiss	france

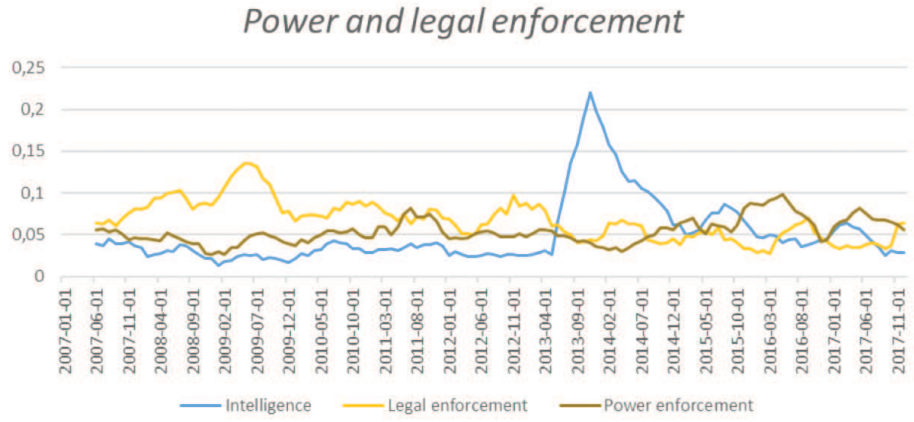

Figure 5. Share of words that were assigned to the topics of *intelligence, legal enforcement* and *power enforcement* in the LDA process, relative to the generated subcorpus (moving average of 6 months).

In the analyzed articles, we are observing a weaponization of data, which became a focus of media coverage mainly because of the Snowden leaks. But in other contexts, too, data is now being used for violent purposes (SZ, 23.3.2015):

> "IS publishes death list with US soldiers
> On a webpage, the jihadists published names, photos and personal data from 100 persons to whom they accused of being involved in the fight against IS in Iraq, Syria, Yemen and other countries. Although a hacker department of IS claimed to have stolen the data from "various servers and databases" of the government..."

While reporting on *legal enforcement* is decreasing (similarly to corporate violations of norms, see above) and reporting on *power enforcement* is increasing only slightly in the context of data and surveillance, the curve that describes the NSA affair has a sharp peak. It follows the typical phases of an issue-attention cycle, from the key event to the post-problem phase.

Datafication

This is about the process of digitization – i.e. the datafication of various aspects of our lives, such as communication or mobility. Coverage is often meant to educate the public. The readers, conceived of as clueless users of smart devices or naive internet surfers, are made aware of the risks and pitfalls of their increasingly digitized world.

Table 4. The 20 most representative words in each of the associated topics *internet* and *technology* (translated).

Internet	Technology
google	car
facebook	data
user	technology
data	cars
internet	devices
net	app
company	people
apple	smartphone
software	software
microsoft	driver
computer	drive
hackers	manufacturer
information	city
privacy	data
advertising	smart
network	future
users	system
Zuckerberg	cameras
Usefulness	kilometer
social	smartphones

Since 2010, at the latest, there has been the perpetual frame of "data octopuses", all-powerful US corporations that are after our data: SZ, 5 October 2010: "How to protect your data: Ten things you can do to protect your privacy on Facebook…", 12 May 2011: "Facebook leak: user data compromised". In addition to Facebook, Google's Streetview also came under fire in 2010, constituting a new frame, "The self as a commodity" (SZ, 5 June 2010). This discourse slackened over the past few years, while coverage of the technological promise of "intelligent" data technologies has increased, in particular towards the end of the research period: "Intelligent heating" (18 May 2017), "Welcome to Dataville" (22 March 2017), "Intelligent roommates" (13 January 2017). The texts are written in a euphoric tone, any concerns about the improper exploitation of personal data are of minor importance. They are aids to adaptation in the sense **that that** they aim to assist the individual in facing the challenges and using the opportunities of the digital age. Hence, the analysis of the top articles published late in the study period suggests that technological discourse is becoming increasingly affirmative.

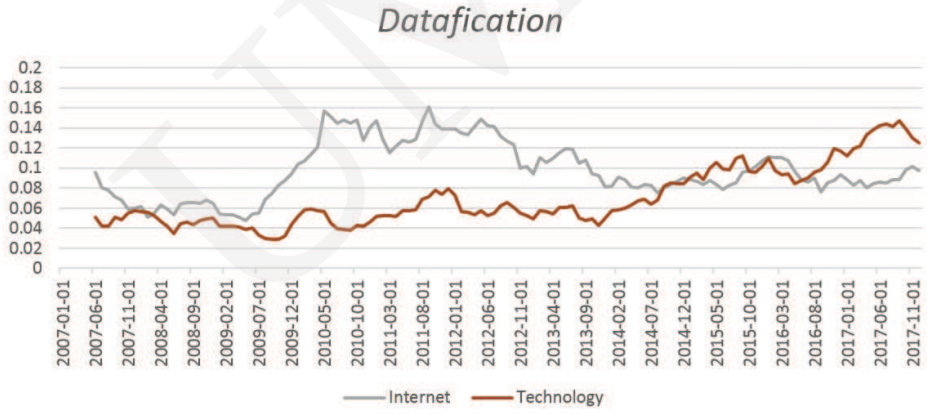

Figure 6. Share of words that were assigned to the topics of *internet* and *technology* in the LDA process, relative to the generated subcorpus (moving average of 6 months).

Conclusion

The research questions presented in the introductory part were: 1) Which facets does news reporting of surveillance, privacy and data security comprise in the *Süddeutsche Zeitung*? 2) Specifically, how are risks to privacy and data security portrayed?

The analysis of articles retrieved from the *Süddeutsche Zeitung* has indeed shown a differentiated picture with regard to the reporting of issues related to surveillance, privacy and data security. Seven meaningful topics could be identified across three different categories. While the first category deals with the *violation of privacy norms*

by the state or private companies, the second category emphasizes the potential of surveillance for *power and law enforcement*, be it legitimate or illegitimate by democratic standards. The third category, *datafication*, deals with the consequences of our ever-more data-driven world, ranging from privacy and data security-related risks to benefits like improved every-day applications or transport, to name but a few. This clearly shows that the much-debated privacy vs. security trade-off is just one, albeit important, aspect of the thematic area analyzed in this article. The *Süddeutsche Zeitung* has also reported from an economic as well as from a technological and a user-oriented perspective, with a variety of connotations ranging from critical to euphoric, from politically loaded to educational.

Based on these findings, how could the mediatized risk culture as portrayed in the *Süddeutsche Zeitung* between 2007 and 2017, be described? First and foremost, it is conspicuous that despite a visible impact of the Snowden leaks, the paper's attention to the violation of privacy norms has been declining significantly over the years. The technological discourse has meanwhile become more affirmative. Given that the Snowden revelations have been conceived of as a game-changer in the debate about surveillance, privacy and data security, this is quite an astonishing finding. We assume this reflects the increasing normalization of surveillance and data harvesting at the cost of individual privacy and data security (see Möller 2017). We consider this finding to be significant since it was obtained from the analysis of a center-left newspaper that often emphasizes the importance of civil rights, and given that German society is traditionally sensitive to privacy and data security-related issues.

This finding is supported by a variety of studies concerning digital issues and the role of industry in shaping public debate. Gillespie (2010), for instance, argued that platforms such as YouTube strategically position themselves regarding how their role in the information landscape should be understood, namely as an opportunity for social and cultural development. The Reuters Institute (2018) found that news media heavily rely on industry sources when reporting artificial intelligence, leaving the debate to self-interested commercial actors instead of making it an object of public concern. The same seems to be the case in the context of privacy and data security, as evident in the case of media coverage about encryption techniques (Thorsen 2016). In the reports we analyzed, the topics related to datafication seem to be less politicized compared to cases where state surveillance was involved, which are much more sensitive due to Germany's historical experiences with two dictatorships. In other words, the potential for scandalization seems to be smaller when risks to privacy and data security are caused by private actors.

Concerning the normative framework for risk communication, the discursive arena model, it is important to note that common users appear to be in a rather passive role, for instance as victims of (legally/morally questionable) surveillance or as naïve individuals who have to be educated about the datafication of our lives. We did not find evidence that the viewpoints of civil society – e.g. in the shape of anti-surveil-

lance NGOs – played a substantial role in the mediated political debate. If at all, the journalists themselves tried to assume the perspective of 'the common man'. Measured against the discursive arena model of risk communication, which postulates a more participatory approach towards risk communication, this could be a shortcoming of *Süddeutsche Zeitung*'s reporting. Another one is the typical event-orientation of news coverage about risk, as we have seen in the context of the declining attention to privacy violations, for instance. When we look at the broad picture of ten years of reporting surveillance, privacy, and data security in the *Süddeutsche Zeitung*, however, it must be said that the newspaper did provide a substantial amount of reporting that could potentially contribute to the sensitization of readers, which is one of the crucial normative expectations of the discursive arena model.

With regard to the generalizability of the results, we ask the reader to bear in mind we conducted an exploratory case study. Additional qualitative content analyses are needed to validate and add more context to our findings. For instance, we find there is need to validate our impression that the voices of civil society did not play a significant role in the analyzed newspaper's reporting. It is in fact a general limitation of the applied method that we cannot answer which sources are used, which actors are given a voice, et cetera. For this purpose, an analysis of the most representative articles on a given topic yields a very accurate conclusion about the latent cognitive pattern that is described by the word distribution of the topic.

The exploratory approach has allowed us to create a category system that could prove useful for further research. We suggest testing and further developing the categories by including more media outlets in the analysis. It could also prove insightful to extend the scope of research across different platforms: This prospect is, for instance, supported by a study that found different framings of whistleblower Edward Snowden in journalistic and social media (Qin, 2015). Given that risk is a highly culture-dependent concept (Beck 2009; Heath, Palenchar 2016), cross-national comparisons could help to identify diverse "risk cultures" concerning the field of surveillance, privacy and data security. However, the latter is surely not the only possible area of research in this context. Internet and data governance, for instance, is a related subject that is significant as political processes are increasingly influenced by the (mis)use of data. However, it is important to keep in mind that security on the internet is not only a policy problem but a societal challenge that requires both public and private actors as well as civil society to take action (European Commission 2013). Therefore, future research should include a variety of perspectives on public communication about cyber risk.

References

Albers M. (2017). Informationelle Selbstbestimmung als vielschichtiges Bündel von Rechtsbindungen und Rechtspositionen [Informational self-determination as a complex bundle of binding regulations and legal positions]. In M. Friedewald, J. Lamla, & A. Roßnagel (Eds.), *Informationelle Selbstbestimmung im digitalen Wandel,* Springer: Wiesbaden, pp. 11–36.

Beck K., Vowe G. (1995). Multimedia aus Sicht der Medien: Argumentationsmuster und Sichtweisen in der medialen Konstruktion [Multimedia from a media point of view. Argumentation patterns and perspectives in the medial construction]. *Rundfunk & Fernsehen*, vol. 43, pp. 549–562.

Beck U. (1992). *Risk Society: Towards a New Modernity*. Sage: New Delhi.

Beck U. (2009). *World at Risk*. Polity Press: Cambridge, UK, Malden.

Blei D. M., Ng A. Y., Jordan M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*. vol. (3), pp. 993–1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. 15.09.2018.

Bonfadelli H. (2004). *Medienwirkungsforschung II: Anwendungen in Politik, Wirtschaft und Kultur [Media effect research II. Applications in politics, economy, and culture]*. UVK: Konstanz.

Brennen J. S., Howard P. N., Nielsen R. K. (2018). An Industry-Led Debate: How UK Media Cover Artificial Intelligence. Retrieved from https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018–12/Brennen_UK_Media_Coverage_of_AI_FINAL.pdf. 15.09.2018.

BSI. (2018). *Glossar der Cyber-Sicherheit [Glossary of Cyber Security]*. Retrieved from https://www.bsi.bund.de/DE/Themen/Cyber-Sicherheit/Empfehlungen/cyberglossar/Functions/glossar.html;jsessionid=BF3C3B1F3AE6F3BE4FDDF7BCB89149CA.1_cid360?cms_lv2=9817278. 15.09.2018.

Budak J., Rajh E., Recher V. (2017). Citizens' privacy concerns. Does national culture matter? In M. Friedewald, J. P. Burgess, J. Čas, R. Bellanova & W. Peissl (Eds.), *Surveillance, Privacy and Security. Citizens' Perspectives* (PRIO new security studies). Routledge: London, pp. 36–51.

Čas J., Bellanova R., Burgess J. P., Friedewald M., Peissl W. (2017). Introduction. Surveillance, Privacy and Security. In M. Friedewald, J. P. Burgess, J. Čas, R. Bellanova & W. Peissl (Eds.), *Surveillance, Privacy and Security. Citizens' Perspectives* (PRIO new security studies). Routledge: London, pp. 1–12.

Chang J., Boyd-Graber J., Gerrish S., Wang C., Blei D. (2009). *Reading tea leaves. How humans interpret topic models. Paper presented at the Neural Information Processing System 2009.*

Deibert R. (2016). Cyber-Security. In M. Dunn Cavelty & T. Balzacq (Eds.), *Routledge Handbook of Security Studies.* Routledge: Abingdon, Oxon, pp. 172–182.

Dencik L., Cable J. (2017). The Advent of Surveillance Realism. Public Opinion and Activist Responses to the Snowden Leaks. *International Journal of Communication,* vo. 11, pp. 763–781.

DiMaggio P., Nag M., Blei D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture. Application to newspaper coverage of U.S. government arts funding. *Poetics,* vol. 41 (6), pp. 570–606.

Dimmroth K., Steiger S., Schünemann W. J. (2017). Outrage without Consequences? Post-Snowden Discourses and Governmental Practice in Germany. *Media and Communication,* vol. 5 (1), pp. 7–16.

Entman R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, vol. 43, pp. 51–58.

European Commission. (2013). *Cybersecurity Strategy of the European Union: An Open, Safe and Secure Cyberspace.* Retrieved from http://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52013JC0001. 15.09.2018.

Gillespie T. (2010). The politics of 'platforms'. *New Media & Society*, vol. 12, pp. 347–364.
Heath R. L., Palenchar M. J. (2016). Paradigms of Risk and Crisis Communication in the Twenty-first Century. In A. Schwarz, M. W. Seeger & C. Auer (Hrsg.), *The Handbook of International Crisis Communication Research.* Wiley: Chichester, UK, pp. 437–446.
Henn P., Vowe G. (2015). Facetten von Sicherheit und Unsicherheit. Welches Bild von Terrorismus, Kriminalität und Katastrophen zeigen die Medien? [Facets of security and insecurity. What picture of terrorism, crime, and disasters do the media show?]. *Medien & Kommunikationswissenschaft,* vol. 63 (3), pp. 341–362.
Jacobi C., van Atteveldt W. Welbers K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism,* vol. 4 (1), pp. 89–106.
Jarvis J. (2011). *The German privacy paradox,* BuzzMachine. Retrieved from https://buzzmachine.com/2010/02/11/the-german-privacy-paradox/. 15.09.2018.
Johnson C. N. (2016). A "Massive and Unprecedented Intrusion". A comparative analysis of American journalistic discourse surrounding three government surveillance scandals. *Digital Journalism,* vol. 5 (3), pp. 318–333.
Keller R. (2003). Distanziertes Mitleiden. Katastrophische Ereignisse, Massenmedien und kulturelle Transformation [Distanced sympathy: Catastrophic events, mass media, and cultural transformation]. *Berliner Journal für Soziologie,* vol. (3), pp. 395–414.
Koppers L., Rieger J., Boczek, K., von Nordheim G. (2018). *tosca (Tools for Statistical Content Analysis)*. Retrieved from https://cran.r-project.org/web/packages/tosca/index.html. 15.09.2018.
Löblich M., Karppinen K. (2014). Guiding Principles for Internet Policy. A Comparison of Media Coverage in Four Western Countries. *The Information Society,* vol. 30 (1), pp. 45–59.
Lüber K. (2014). *Data privacy made in Germany,* Goethe-Institut. Retrieved from https://www.goethe.de/en/m/kul/med/20446236.html. 15.09.2018.
Maier D., Waldherr A., Miltner P., Wiedemann G., Niekler A., Keinert A. (2018). Applying LDA topic modeling in communication research. Toward a valid and reliable methodology. *Communication Methods and Measures,* vol. 54 (10), pp. 1–26.
Möller J. (2017). Farewell to an utopia. Technology discourse in the German NSA debate. In S. Tosoni, N. Carpentier, M. F. Murru, R. Kilborn, R. Kunelius, A. McNicholas et al. (Eds.), *Present scenarios of media production and engagement.* Edition Lumière: Bremen, pp. 173–184.
Nordheim G. von, Boczek K., Koppers L. (2018). Sourcing the Sources. *Digital Journalism,* vol. 6 (7), pp. 807–828.
Oggolder C. (2012). Inside – outside: web history and the ambivalent relationship between old and new media. *Historical Social Research,* vol. 37 (4), pp. 134–149.
Oggolder C. (2015). From Virtual to Social. Transforming Concepts and Images of the Internet. *Information & Culture,* vol. 50 (2), 181–196.
Pantti M., Wahl-Jorgensen K., Cottle, S. (2012). *Disasters and the Media*. Peter Lang: New York.
Patnode R. (2003). Of Viruses and Victims: Framing the Internet, 1988–1990. In ICA (Eds.), *Conference Proceedings*, pp. 1–20.
Puschmann C., Scheffler T. (2016). Topic modeling for media and communication research. A short primer. *HIIG Discussion Paper Series* (5), pp. 1–17. Retrieved from https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2837801_code1779785.pdf?abstractid=2836478&mirid=1. 15.09.2018.
Qin J. (2015). Hero on Twitter, Traitor on News. How Social Media and Legacy News Frame Snowden. *International Journal of Press/Politics,* vol. 20 (2), pp. 166–184.
Rauchenzauner E. (2008). *Schlüsselereignisse in der Medienberichterstattung [Key Events in Media Coverage]*. Wiesbaden.

Roslyng M. M., Eskjær M. F. (2017). Mediatised risk culture: News coverage of risk technologies. *Health, Risk & Society,* vol. 19 (3–4), pp. 112–129.

Rössler P. (2001). Between online heaven and cyberhell. The framing of 'the internet' by traditional media coverage in Germany. *New Media & Society,* vol. 3 (1), pp. 49–66.

Schäfer M. S., Ivanova A., Schmidt A. (2011). Global Climate Change, Global Public Sphere? Media Attention for Climate Change in 23 Countries. *Studies in Communication and Media,* vol. 1, pp. 133–148.

Sklar A. (1997). (De)constructing the (Information) Highway. Discourse Analysis of Canadian Popular Press. *The Electronic Journal of Communication,* vol. 7 (4). Retrieved from http://www.cios.org/EJCPUBLIC/007/4/007414.HTML. 15.09.2018.

Thorsen E. (2016). Cryptic Journalism. News Reporting of Encryption. *Digital Journalism,* vol. 5 (3), pp. 299–317.

Wäscher T. (2016). Framing Resistance Against Surveillance. Political communication of privacy advocacy groups in the "Stop Watching Us" and "The Day We Fight Back" campaigns. *Digital Journalism,* vol. 5 (3), pp. 368–385.

Zeller F., Wolling J., Porten-Cheé P. (2010). Framing 0/1. Wie die Medien über die "Digitalisierung der Gesellschaft" berichten [Framing 0/1. How media report on the "digitization" of society]. *Medien & Kommunikationswissenschaft,* vol. 58 (4), pp. 503–524.

Zellig S. H. (1954). Distributional Structure. *WORD,* vol. 10 (2–3), pp. 146–162.