

RESEARCH NOTE

Measuring the Public Agenda using Search Engine Queries

Michael Scharkow and Jens Vogelgesang

Institut für Kommunikationswissenschaft, University of Hohenheim, Stuttgart, Germany

If you type Google into Google, you can break the Internet.
(*Jen Barber, Head of IT in the British Comedy "The IT-Crowd"*)

The research note we present here explores the idea of measuring the public agenda without employing classical survey techniques. Drawing on data from the web service Google Insights for Search (GIFS), we evaluate the use of online observation to measure information seeking. Hereby, we follow the ideas of McLeod, Becker, and Byrnes (1974) that classical first-level agenda setting has to be both theoretically and empirically enriched by considering subsequent behavioral audience responses stemming from prior saliences. Using the terminology proposed by Becker, McCombs, and McLeod (1975, p. 39), we distinguish between “first-level” and “subsequent” audience responses. While the former refers to cognitive concepts such as awareness or salience, the latter typically refers to behavioral concepts such as candidate choice or turnout at the polls.

In what follows, we focus on what we call immediate subsequent audience responses. For example, people may want to know more about an issue (salience-driven media use) or even want to talk about it (salience-driven conversation). We subsume both salience-driven behaviors under the more general concept of information seeking. We assume that immediate audience responses as such only emerge when a certain degree of issue salience has been reached. In addition, other recipient characteristics like being politically interested or having need for orientation (Huck, Quiring, & Brosius, 2009) may also foster immediate audience responses.

The theoretical distinction between first-level and subsequent audience responses made by Becker et al. (1975) has a major methodological implication. First-level audience responses are cognitive, and thus unobservable in nature. In contrast, immediate audience responses manifest in everyday life and can be directly observed by any researcher.

Methodologically speaking, salience represents an unobservable latent variable, whereas immediate audience responses such as salience-driven media use represent observable manifest variables. When immediate audience responses emerge from issue salience, there should be a relationship between the two concepts. Consequently, if there is a strong correlation between online information seeking and issue salience, the former could be used as a convenient proxy for the latter.

In this research note, we investigate how Google search inquiries relate to traditional measures of the public agenda. First, we briefly recapitulate how to measure the public agenda by means of public opinion polls. Second, we introduce GIFS as an alternative observational tool to measure the public agenda. Third, we present a case study and compare GIFS results with public agenda survey data. Fourth, we conclude by discussing the potential merits and drawbacks of using GIFS for further agenda setting research.

The Measurement of the Public Agenda

Measuring the Cognitive Dimension using Surveys

The reliable and valid measurement of issue salience is a persistent problem of agenda-setting research. Issue salience is typically measured by means of special questions in public opinion polls (Dearing & Rogers, 1996, p. 45–49) focusing on facets like perceived issue salience or interpersonal issue salience of the “multi-faceted concept” (Gadziala & Becker, 1983, p. 122) termed salience. However, public opinion polls can only provide indirect measures of the public agenda, since agenda-setting researchers depend on what people report about the most important problem facing the country. Even with the most sophisticated designs that combine data from closed- and open-ended survey questions, researchers do have to rely on the self-reports given by the respondents. This is equally true for studies in which behavioral aspects of agenda setting, such as interpersonal conversations, were measured using survey responses (for an early example, see Atwood, Sohn, & Sohn, 1976). Put generally, the problem of how to ask about issue salience (or one of its facets) remains to this day (cf. Edelstein, 1993; McLeod et al., 1974; Niemi & Bartels, 1985). Because of that we subsequently explore the utility of GIFS as an alternative observational, non-obtrusive tool to measure the public agenda.

Measuring the Behavioral Consequences using (Online) Observation

The behavioral consequences of issue awareness and salience, like immediate information seeking or follow-up conversations are basically observable processes. Unlike in journalism studies (Quandt, 2008) or media psychology (Ravaja, 2004; Unz & Schwab, 2005), observation is rarely used in mass communication research. This is most obviously due to the difficulty of observing more than a handful of selected people at the same time outside of a laboratory. Consequently, agenda setting field studies employing participant observation are quite limited in scope (Kepplinger & Martin, 1986).

A possible solution to the small sample problem of field studies employing participant observation lies in automated observation tools that work without human observers. The most commonly used method of this kind are audience ratings (Webster, Phalen, & Lichty, 2000), which enable researchers to observe the media use of many recipients in real time. However, these data are less well-suited for agenda-setting research because: (a) interest can only be measured at the granularity of programmes not issues; (b) the users' possibilities for salience-driven information seeking are limited in linear media like TV; and (c) information seeking may not be the only relevant motive for programme choice.

A relatively new and, we argue, promising way of unobtrusive observation has been available since the advent of the internet, both for observing follow-up communication and information seeking. Any communication that happens via protocols like HTTP (World Wide Web), SMTP (Email), or NNTP (Usenet) can be logged, stored, and analyzed with some technical but little human effort. For example, Roberts, Wanta, and Dzwo (2002) examined follow-up communication in electronic bulletin boards (EBB). The authors analyzed the content of four media outlets as well as AOL's political message board. For three out of four selected issues (immigration, health care, taxes, and abortion) significant positive cross-correlations between the media content and the online discussions were found, leading the authors to conclude that "[m]edia coverage apparently can provide individuals with information to use in their Internet discussions" (Roberts et al., 2002, p. 464). While the analysis of user-generated content, like message boards, comments, chats, or Twitter messages, promises to be a powerful method for the measurement of follow-up communication (Shamma, Kennedy, & Churchill, 2009), we are interested here in observing information-seeking behavior. Fortunately, the Internet provides users with a tool for this task—search engines which are arguably the first, and possibly only, device for information seeking for millions of people. According to recent estimations by comscore, U.S. online users submitted 137 billion search queries in 2008, which means 1.7 queries per day and user (South Florida Business Journal, 2009).

Following Bloj (1975, as cited by Becker et al., 1975), we argue that aggregate search queries as subsequent behavioral effects of salience provide valid and reliable measures of the public agenda. Every individual search query requires at least awareness of an issue as well as a certain degree of commitment. For a political issue this individual commitment may plausibly arise from its salience. Of course, people may also employ online search engines for several other reasons, such as interest or curiosity. But even if search queries for political issues do not originate from salience alone, we assume that over time variability in aggregate numbers indicates changes in the public agenda.

From a measurement point of view, search queries have many desirable properties: Compared to survey questions, there is no interviewer bias or social desirability involved, the measurement is completely unobtrusive and happens in the field. Moreover, for many users, there is virtually no effort involved in using search engines, compared to buying a book or searching a paper encyclopedia. Of course, the demographic of internet users is still different from the general population, so that we cannot take the complete public agenda, as measured by search queries, at face value. In case of Germany, the ARD/ZDF-Online study 2009 shows that 43.5 million Germans (67.1%) use the internet at least occasionally (Eimeren & Frees, 2009). Comparing the

demographics of the German online versus offline population reveals that those without internet access are predominantly female, age 60 years and over and less educated (Gerhards & Mende, 2009). But unless the underlying mechanisms of seeking information about salient issues are fundamentally different for onliners than for offliners, the validity of online search queries for measuring the public agenda is not at risk.

Until recently, there has been the lack of availability of search query data for research purposes. Basically, only the providers of search engines like Google, Yahoo, or Microsoft are able to collect and store such data which are heavily used by the companies themselves in order to optimize services for users and advertising partners. Applications of scientific analysis of those queries were therefore quite rare but prominent, possibly because it could be demonstrated that searches for pornographic and illegal downloads accounted for the most frequent queries (Silverstein, Marais, Henzinger, & Moricz, 1999).

Fortunately, the world's largest search engine, Google, has recently begun to make aggregate log file results available to the public. Google started their web service Insights for Search (<http://google.com/insights/search>) in August 2008 as a follow-up to their earlier Google Trends site. GIFS provides public access to Google's logged search queries and allows for filtering by search term, time frame, and region. Unlike older services, users cannot only see graphical presentations of these data, but also download actual data tables of the search volume for a particular query. In the remainder of this research note, we will investigate the utility of this data for agenda-setting research—case study.

Method

In order to check the validity of Google search queries for the measurement of the public agenda, we compare the aggregate search query data provided by GIFS with aggregate survey data from telephone interviews conducted by the FORSA institute. According to the arguments presented above, search queries can be seen as behavioral consequences of issue salience. If the hypothesized relationship is true, we expect to see a strong correlation between the two time series.

As a case study, we use a single-issue study on the German General Election 2005. Specifically, we focus on Paul Kirchhof, a former judge at the Federal Constitutional Court, Professor at the University of Heidelberg, and fiscal expert in Angela Merkel's campaign team. Kirchhof's controversial ideas on a flat income tax as well as later comments on social issues and gender roles generated much media attention and debate, leading to his withdrawal from the shadow cabinet just after the election (Schmitt-Beck & Faas, 2006).

We chose the issue Paul Kirchhof for this case study mainly for pragmatic reasons. First, Kirchhof appears in public only for a limited period during the campaign, from his nomination in Angela Merkel's government team on August 16th, until some time after election day on September 18th. This enables us to restrict the coding of survey answers and data retrieval from Google to about 6 weeks. Second, a person's name can rather unambiguously be used in search queries, making it easier for the coders of the survey data to find references to this issue. Applying our approach to more complex issues will certainly require more thought and effort.

Search Query Data

The data on aggregate search queries was retrieved from the GIFS website with the following settings (Figure 1). The selection of data was restricted to Germany, as defined by certain ranges of IP numbers from which the search query originated. The time range requested was August/September 2005, from which we removed the first 2 weeks afterwards.

Two important restrictions are currently implemented in GIFS. Daily search volume data can only be retrieved for up to 8 weeks from Google, for longer periods only weekly data are provided which is of less use for most agenda-setting research. This could be remedied by sliding a 2-month window over the period of interest, yielding six results sets for a year. However, the second restriction from GIFS makes merging these data difficult: The search volume is only available as normalized data, based on the relative frequencies of search queries, with the peak day within the period as a reference point. This is less problematic for our time series correlation, but estimates of absolute measures are currently impossible to obtain using only GIFS.

The specification of the query string for GIFS is very simplistic. There is no support for complex queries connected by AND or NOT, no fuzzy matching or wildcards (cf. Hollanders & Vliegenthart, 2008). For many applied research cases, manual specification of different query variants and aggregation of the retrieved data will be necessary. For our case study, a simple search string is easily constructed. In order to account for different spellings of the name we use the following search terms, the plus sign marks a Boolean OR: “kirchhof+Kirchhoff.”

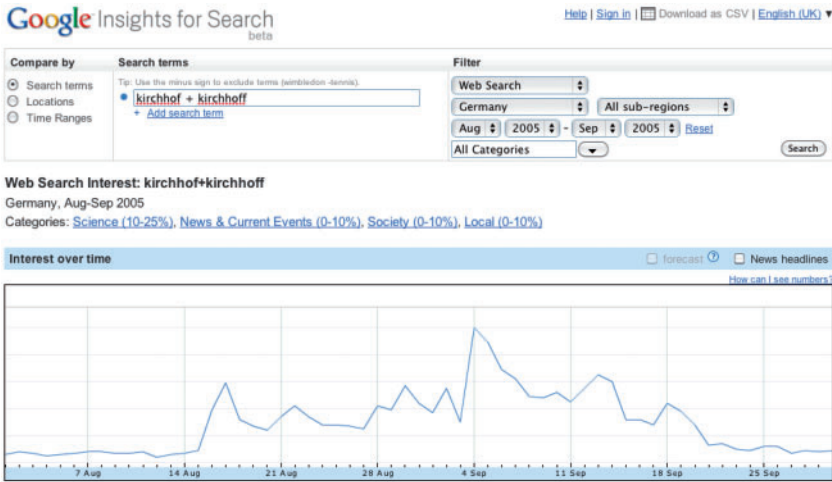
Using this query yields, the result page depicted in Figure 1 which includes a graphical display of the time series, matching results from Google’s news archive, and a downloadable CSV file with the numbers. The latter is only available to logged-in users while the rest of the interface is public.

Survey Data

The survey-based time series in our analysis is the aggregate number of responses to an open-ended question about important issues. The data were collected by FORSA using standardized computer-assisted telephone interviews. The survey covers the population of the eligible voters of Germany aged 18 years and over, the sample was drawn using the German standard CATI sampling procedure (Gabler & Häder, 2002). On average, 500 interviews were conducted every weekday during the period of interest in 2005. Unfortunately, this means that there are no survey data available for the weekends. Daily response rates ranged from 50% to 60% (personal correspondence with FORSA).

The open-ended question asked was: “Can you recall any important issues recently covered in the news media that interest you?” and the responses were entered as free text by the interviewers. The wording of the question is multi-dimensional and does not clearly capture one of the processes described in the introduction. Rather unfortunately, this question measures issue awareness (“recall”), salience (“important issues”), and interest at the same time. We do not expect this to be an issue for correlational analyses but caution against inferring absolute levels of issue salience

Figure 1
Result page of Google Insights for Search



from this question. Two students analyzed all responses and coded every answer referring to Paul Kirchhof. For future analyses with equally simple search terms, a dictionary-based automatic coding is certainly feasible (cf. Krippendorff, 2004).

Results

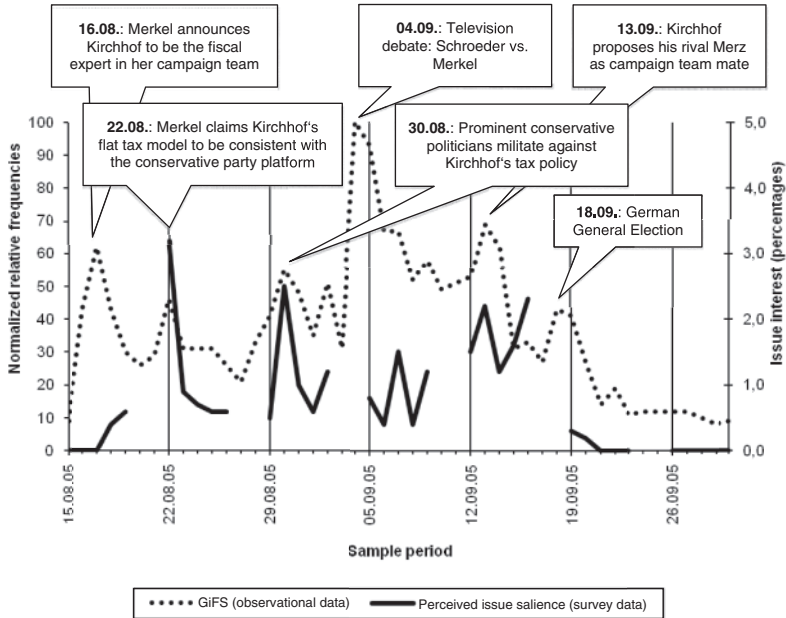
In absolute numbers, only a small fraction of the respondents named Paul Kirchhof an important topic. Kirchhof does not appear on the public agenda until his nomination and then quickly peaks at about 3% of all answers, then a rather volatile level of interest and again a small rise in the week before the election (Figure 2).

The time series for Google search queries related to Paul Kirchhof looks somewhat similar to the aggregate survey responses. There is a sharp rise in interest on September 4th, the day of the first TV debate between Merkel and Schröder, in which Kirchhofs proposals for taxation were heavily debated. Another peak in search queries occurs some 10 days later, after Kirchhof suggested forming a team with prominent CDU fiscal expert Friedrich Merz.

We computed a simple correlation between the two time series ($r = .49$, $p < .01$). First of all, this modest correlation suggests that search engine queries are a somewhat different but still related measure of the public agenda compared with traditional survey measures. However, there are still some caveats to consider: For example, unlike Krause and Fretwurst (2007) we did not specify ARIMA cross-correlations or transfer functions because of too few data points and too much missing survey data due to the weekends. There are indications of an autoregressive pattern within both times series which is hardly surprising (i.e., the aggregate levels of salience and information seeking are dependent on their preceding values). Hence, we expect that some portion of the common variance between the two time series is due to the

Figure 2

Survey responses and search queries compared for Paul Kirchhof



Note: The black line is discontinuous because no survey data are available for weekends.

inherent autoregressive process within each time series. Nonetheless, we conclude that the correlation between two time series collected by different means is a promising indicator that search engine queries represent a somewhat new facet of the public agenda.

Discussion

As Huck et al. (2009, p. 5) recently stated, “the behavioral consequences of agenda setting [...] are still largely unexplored”. In this research note, we introduced a novel way of measuring the public agenda by using aggregate data from queries to the Google search engine. Although the presented results are based on a rather simple case study, we are confident that using Google Insights for Search provides a promising tool for the study of agenda-setting processes. First of all, the data collection using GIFS is free of cost, easy to understand and very quick, even when some tuning of the search query is required. The availability of daily search volume data without missing data makes time series analysis more straightforward.

There are, however, some challenges and open questions that need to be addressed, both methodologically and empirically:

- 1. We do not know yet how many people turn to Google for a given issue, and how this relates to the number of respondents in telephone surveys.

The absolute amount of Google search queries related to, for example, Paul Kirchhof is not (yet) publicly available. However, this is less of a problem for agenda setting research using time-series correlations.

2. Since Google only provides the aggregate data as-is, we cannot know or test how exactly queries are counted. It remains unclear, for example, whether the search volume for a term is increased by repeated searches of the same individual.
3. On a related issue, the normalization of GIFS data makes comparisons between countries, regions, and time spans difficult. Further refinement in dealing with data provided by Google will be necessary.
4. Since the survey respondents and Google users are certainly from different populations, it is yet unclear whether the observed correlation can easily be generalized. After all, Internet users are at least younger and better educated than the general population. This becomes even more important when actual agenda-setting processes relating media content with the public agenda are investigated online (cf. Roberts et al., 2002). However, within 10 years, the generational gap in Internet use is likely to be much smaller than today.
5. As with conventional aggregate measures in agenda-setting research, the problem of an ecological fallacy persists when using GIFS data (Robinson, 1950). Put differently, the aforementioned correlation is based on aggregate data. This correlation based on aggregate data is likely to differ from a correlation based on individual-level data (Erbring, 1990). The caution to be aware of is that variation in the GIFS and survey times series cannot simply be attributed to individual changes in salience.

We acknowledge the need for further theoretical elaboration and empirical research on immediate audience responses. Specifically, the relationship between awareness, interest and salience and their behavioral consequences (i.e., media use and conversation seeking) needs further exploration on the individual rather than the aggregate level. As a reviewer of this research note rightfully pointed out, the relationship between salience and information seeking itself may be dynamic as well as issue- and person-specific. The general conditions under which GIFS data are (not) valid proxy variables for issue salience still needs further examination.

Nonetheless, since GIFS data are available for virtually all imaginable topics, for many countries and even regions, and go back for several years, a plethora of interesting empirical research projects may profit from this data source. This includes analyses of media resonance to public relations and advertising campaigns, cross-country and over-time comparisons of public interest in political and social issues, even forecasting tomorrow's news from today's recipients' information seeking. Prospectively, combining GIFS data with media monitoring data (e.g., Google News) may provide new insights into agenda-setting dynamics. This approach is already successfully employed for epidemiological research (Freifeld, Mandl, Reis, & Brownstein, 2008, see also <http://www.google.org/flutrends>).

In sum, the Internet age clearly seems to require further theoretical elaboration and empirical examination of this new form of public agenda dynamics. We look forward to seeing follow-up studies using online data and more complex issues, covering a longer time span, and using more elaborate techniques of times series analysis.

References

- Atwood, L., Sohn, A., & Sohn, H. (1976). Community discussion and newspaper content. Paper presented at the Annual Convention of the Association for Education in Journalism. College Park, MD: University of Maryland.
- Becker, L., McCombs, M., & McLeod, J. (1975). The development of political cognitions. In S. Chaffee (Ed.), *Political Communication* (pp. 21–63). Beverly Hills: Sage.
- Bløj, A. G. (1975). Into the wild blue yonder: Behavioral implications of agenda-setting for air travel. In M. McCombs & G. Stone (Eds.), *Studies in agenda-setting*. Newhouse Communication Research Center, Syracuse University.
- Dearing, J., & Rogers, E. (1996). *Agenda-setting*. Thousand Oaks, CA: Sage.
- Edelstein, A. (1993). Thinking about the criterion variable in agenda-setting research. *The Journal of Communication*, 43(2), 85–99.
- Eimeren, B. V., & Frees, B. (2009). Der Internetnutzer 2009 – multimedial und total vernetzt? [The Internet User 2009 – multimedial and completely linked-up?]. *Media Perspektiven*, (7), 334–348.
- Erbring, L. (1990). Individuals writ large: An epilogue on the ecological fallacy. *Political Analysis*, 1, 235–269.
- Freifeld, C. C., Mandl, K. D., Reis, B. Y., & Brownstein, J. S. (2008). HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *Journal of the American Medical Informatics Association*, 15(2), 150–157.
- Gabler, S., & Häder, S. (2002). *Telefonstichproben. Methodische Innovationen und Anwendungen in Deutschland*. [Telephone samples. Methodological innovations and applications in Germany]. Münster: Waxmann.
- Gadziala, S., & Becker, L. (1983). A new look at agenda-setting in the 1976 election debates. *Journalism Quarterly*, 60(1), 122–126.
- Gerhards, M., & Mende, A. (2009). Offliner: Ab 60-jährige Frauen bilden die Kerngruppe. [Offliner: Women at sixty and upward form the main group]. *Media Perspektiven*, (7), 365–376.
- Hollanders, D., & Vliegthart, R. (2008). Telling what yesterday's news might be tomorrow: Modeling media dynamics. *Communications*, 33(1), 47–68.
- Huck, I., Quiring, O., & Brosius, H. (2009). Perceptual phenomena in the agenda setting process. *International Journal of Public Opinion Research*, 21(2), 139–164.
- Kepplinger, H., & Martin, V. (1986). Die Funktion der Massenmedien in der Alltagskommunikation. [The function of media in everyday communication]. *Publizistik*, 31(1–2), 118–128.
- Krause, B., & Fretwurst, B. (2007). Kurzfristige Agenda-Setting-Effekte von Fernsehnachrichten. Eine Zeitreihenanalyse am Beispiel Ausländerfeindlichkeit und Rechtsradikalismus [Short-term agenda setting effects of television news.]. In B. Krause, B. Fretwurst & J. Vogelgesang (Eds.), *Fortschritte der politischen Kommunikationsforschung: Festschrift für Lutz Erbring* (pp. 171–196.). Wiesbaden: VS Verlag für Sozialwissenschaften.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.

- McLeod, J., Becker, L., & Byrnes, J. (1974). Another look at the agenda-setting function of the press. *Communication Research*, 1(2), 131–166.
- Niemi, R., & Bartels, L. (1985). New measures of issue salience: An evaluation. *Journal of Politics*, 47(4), 1212–1220.
- Quandt, T. (2008). Methods of journalism research: Observation. In M. Löffelholz, D. Weaver & A. Schwarz (Eds.), *Global journalism research: Theories, methods, findings, future* (pp. 131–142). Malden, Oxford, Carlton: Blackwell.
- Ravaja, N. (2004). Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology*, 6(2), 193–235.
- Roberts, M., Wanta, W., & Dzwo, T. (2002). Agenda setting and issue salience online. *Communication Research*, 29(4), 452–465.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–357.
- Schmitt-Beck, R., & Faas, T. (2006). The campaign and its dynamics at the 2005 German General Election. *German Politics*, 15(4), 393–413.
- Shamma, D., Kennedy, L., & Churchill, E. (2009). Tweet the debates: Understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media* (pp. 3–10). New York, NY: ACM.
- Silverstein, C., Marais, H., Henzinger, M., & Moricz, M. (1999). Analysis of a very large web search engine query log. *Proceedings of the ACM SIGIR Forum* (pp. 6–12). New York, NY: ACM.
- Comscore: Online traffic soars, sales growth slows. (January 30, 2009). *South Florida Business Journal*. Retrieved from <http://southflorida.bizjournals.com/southflorida/stories/2009/01/26/daily48.html>.
- Unz, D., & Schwab, F. (2005). Viewers viewed: Facial expression patterns while watching TV news. In L. Anolli, S. Duncan, M. Magnusson & G. Riva (Eds.), *The hidden structure of social interaction. From genomics to cultural patterns* (pp. 253–263). Amsterdam: IOS Press.
- Webster, J., Phalen, P., & Lichty, L. (2000). *Ratings analysis: The theory and practice of audience research*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Wlezien, C. (2005). On the salience of political issues: The problem with ‘most important problem’. *Electoral Studies*, 24(4), 555–579.

Biographical Notes

Michael Scharkow is research associate at the University of Hohenheim, Germany. His research interests include empirical research methods, online communication and media use.

Jens Vogelgesang is research associate at the University of Hohenheim, Germany. His research fields include media effects and migration, European public sphere, and complex statistical methods.