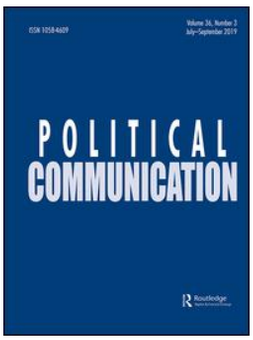

Political Communication

ISSN: 1058-4609 (Print) 1091-7675 (Online) Journal homepage: https://www.tandfonline.com/loi/upcp20

The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research across National Jurisdictions Using Sensitive Data

Wouter Van Atteveldt, Scott Althaus & Hartmut Wessler

To cite this article: Wouter Van Atteveldt, Scott Althaus & Hartmut Wessler (2020): The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research across National Jurisdictions Using Sensitive Data, Political Communication, DOI: 10.1080/10584609.2020.1744780

To link to this article: https://doi.org/10.1080/10584609.2020.1744780

Published online: 18 May 2020.

POLITICAL COMMUNICATION
https://doi.org/10.1080/10584609.2020.1744780

Routledge
Taylor & Francis Group

The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research across National Jurisdictions Using Sensitive Data

Wouter Van Atteveldt[a], Scott Althaus[b], and Hartmut Wessler[c]

[a]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands; [b]University of Illinois at Urbana-Champaign, Urbana-Champain, USA; [c]University of Mannheim, Mannheim, Germany

ABSTRACT

Open science and effective collaboration both require the sharing of data between researchers. This is especially true for computational methods, as the technical complexity and heterogeneous data sources often require collaboration between researchers in different institutions and jurisdictions. Many data sources, however, cannot be shared openly because of copyright law and contracts such as terms of service. These regulations can be complex, sometimes untested in case law, and vary between countries and over time. This paper details our experiences in conducting international comparative research on very large collections of news items from multiple countries. We set out the main problems we have encountered and some short- term approaches we have used to mitigate some of these problems. We end with listing some additional long-term actions that will advance our research community's ability to collaborate on computational research using sensitive data.

KEYWORDS

open science; copyright; computational social science

Computational social science research is booming, at least looking at the number of recent special issues and edited volumes in political science, communication science, and related fields (Domahidi et al., 2019; Shah et al., 2015; Van Atteveldt & Peng, 2018). Given the complexity of many computational projects and the difficulty of finding all required skills in a single group, computational research often requires collaboration between different groups. Especially in comparative research based on text analysis, the linguistic skills, algorithmic tools, and contextual expertise needed to analyze political text from different countries often requires a collaboration between at least a subset of the countries involved.

Existing projects using public data such as the Comparative Manifesto Project and Comparative Agendas Project show that such collaboration is feasible and potentially very productive. The challenge for political communication research is that the raw texts are often not public: media content is generally copyrighted, social media and other private speech can be privacy-sensitive, and data acquired from third parties often cannot be shared or published under terms of use. This creates problems for collaboration when different teams are needed to collaborate on conducting the text analysis. This also makes it difficult to share or publish data under Open Science Principles.

This paper aims to share some of the experiences and lessons learned over the past years of conducting large scale multilingual text analysis involving collaborators in different physical locations.The authors are senior researchers from the US and two

CONTACT Wouter Van Atteveldt w.h.van.atteveldt@vu.nl De Boelelaan 1101, 1081 HV Amsterdam, The Netherlands.

different European countries who have been working together for several years in a longitudinal and comparative automatic text analysis project using a variety of non-public data sources. Moreover, the authors have conducted various other collaborative text analysis projects using sensitive data and worked toward creating text analysis tools and setting up large scale databases of political texts. We offer what follows in the hope that others can benefit from our several combined years of frustrating experiences, frequent mistakes, and small victories in the realm of cross-national collaborative text-analytic work using sensitive data.

Open Science Requires Open Data, But Sensitive Data Isn't Easily Shared

Collaboration on computational research, even more than on traditional research, requires the sharing of both data sets and the tools and scripts used to process these data. In larger research consortia, the technical, theoretical, and local expertise to conduct specific analyses are often distributed among different teams. Tools developed in one team need to be adapted and validated for different contexts, often requiring both linguistic and substantive expertise from other teams. In particular, cross-national comparative studies generally require close collaboration between teams that may be located in different national jurisdictions. In these cases, being able to freely share materials is crucial for efficient collaboration and for ensuring the validity of measurements.

Sharing research materials is also a crucial part of open science (Klein et al., 2018; Miguel et al., 2014; Nosek et al., 2015). Data transparency is a key part of the move toward transparent and open science (Nosek et al., 2015), which improve the reproducibility and robustness of scientific findings by allowing other scholars to inspect and verify published results (Klein et al., 2018). Moreover, sharing data can improve the efficiency of science by allowing greater re-use and more collaboration and specialization (Van Atteveldt et al., 2019).

In many cases, however, researchers are not free to share sensitive data, which for the context of this discussion we will define as data that originates from third parties and that cannot be openly shared due to legal, proprietary, or regulatory barriers. There are many types of sensitive data that might be of interest to political communication researchers, such as social media data and survey or experimental data identifying individuals, which can be covered by privacy regulations such as the EU General Data Protection Regulation (GDPR). However, given the scope of this contribution we will focus here on sharing politically-relevant media content, such as entire newspaper articles or complete transcripts of television news broadcasts.

There is an important caveat that must be underscored for what follows: none of the authors has legal expertise, and our understanding of the relevant legal landscape may be partial or flawed. We offer no legal advice here, but merely convey our imperfect understanding of legal barriers that define boundaries we have been working to uphold while still advancing research projects using sensitive data.

Barriers to Sharing Political Media Content

There are at least three factors hindering the sharing of full text political media content: copyright laws; contract laws/terms of service; and how these regulations vary between countries and over time.

Copyright law is a temporary monopoly on the distribution of text and other creative works intended to allow authors to make money from their creations. The copyright on most media content is owned by the company that owns the media outlet. Contract laws and terms of service come into play when media content holdings are obtained from library sources and from commercial content database providers such as LexisNexis or Factiva, which offer media content under general campus licenses or other contractual agreements. Contract laws and terms of service also come into play when researchers scrape media content holdings from Internet sites or when researchers enter into formal agreements with media content owners. It is also important to emphasize that even when researchers located within a particular country operate within that country's established copyright laws, contract laws and terms of service might still limit their ability to share (or even use) the news content that they have access to. When there is a conflict between the usage terms imposed by contract and by copyright law, in many cases it is not clear which set of laws should prevail. For example, a researcher in the United States might use LexisNexis news data within "fair use" exemptions in copyright law, but still be in violation of the campus contract that allowed the researcher access to LexisNexis in the first place. In addition, sensitive material that has been acquired by one project team cannot in most cases be physically transferred across campus or national boundaries for use by other teams in a collaborative project. Commercial content providers might require identical licenses held by the collaborating institutions for material to be used in more than one place, and in many cases researchers will have no access to or understanding of the terms of the license that their campus is bound by.

Finally, legal boundaries differ across jurisdictions, are constantly evolving, and are often poorly understood by campus authorities. Copyright law and contract law differs not only between the US and EU, but also between EU member states. For example, while the United States has a "fair use" exception to copyright law that is generally favorable to researchers, the concept of "fair use" has no direct parallel in the European Union even though some research uses may be exempted from copyright restrictions. Moreover, legal barriers that govern sharing of sensitive data are constantly evolving (e.g., the Digital Millennium Copyright Act and the new EU Copyright Directive). While US copyright law has broader fair use exemptions, it also has harsher (statutory) damages; and while the new EU Copyright Directive has specific exemptions for academic use, these provisions are untested and need to be written into national law before taking effect, potentially introducing more variation and uncertainty. On terms of service, there are differences between jurisdictions for example, in whether "click through" agreements or terms of service simply posted on a website constitute a valid contract. This issue was at the heart of the prosecution (and subsequent suicide) of Aaron Schwartz under the US Computer Fraud and Abuse Act for violation the JSTOR terms of service by automatically downloading large amounts of scientific articles from their archive.

The Need for Finding Solutions within Legal Boundaries

The complexities mentioned above can result in two extreme reactions. Many individual researchers and research groups (especially in computer science) simply ignore legal barriers to scrape and use the data they want. However, if researchers make a mistake or get caught in a copyright violation, usually the consequences will fall heavier on their

institutions than on themselves. Institutional risk managers therefore often take the other approach and minimize risk by disallowing any sharing of sensitive data altogether.

Neither approach is satisfactory from a data transparency perspective, however, as even researchers that gather data without permission cannot share these in an open way. Thus, it is important to develop practices that foster research transparency within the legal and ethical bounds set by relevant regulations. Part of the solution consists of things can be done right now by individual research groups, while a fuller solution will depend on long-term advocacy and education efforts by the research community as a whole.

Some compliant solutions to consider for the short term can include:

Publishing or Sharing Small Validation Sets

Depending on the exact data source and terms of service, it might be allowed to publish a small sample of sensitive material to allow for analyses to be checked by others. Although this can be used to check rule-based analyses such as dictionaries, it is less useful for validating or improving corpus analysis and supervised or unsupervised analyses such as scaling or topic modeling because these methods' results vary strongly with the size of the dataset.

Publishing Metadata

In some cases, such as online news or data from Twitter or LexisNexis, other researchers might be able to retrieve the same data used by an originating research team given the identifying metadata such as URL, status ID, or article headline and date. This can be cumbersome and costly, however, if large amounts of data are needed to duplicate or validate analyses. Moreover, the persistence of the remote data can often not be guaranteed, jeopardizing the future reproducibility of research. Archiving an encrypted version of the sensitive data could solve that problem, but will presumably run into the same regulatory hurdles.

Meeting Face to Face

If data cannot cross institutional barriers, the easiest way to collaborate on data can be to physically come together. For example, many campus licenses governing sensitive news data have exceptions for visiting scholars who are physically on campus premises. This is not a solution, however, for sharing data with external parties, if data from multiple institutions need to be analyzed jointly, or if financial or agenda constraints prevent meeting long enough to do substantial analytical work on the data. Growing concerns about the environmental impact of travel within the academic community might also hinder face-to-face meetings when long-distance flights would be required to bring collaborators together.

Remote Access to Computer Systems

Essentially a virtual form of meeting face-to-face, it might be possible to give collaborators remote access to the relevant computer system on which sensitive data are physically stored. Depending on exact agreements governing how particular forms of sensitive data

might be used, this might overcome some legal problems of sharing data within trusted collaborations. It is generally difficult to prevent remote users from downloading the data, however, so this might pose risks to the institution and does not solve the problem of sharing the data outside trusted collaborations.

Non-consumptive Research

One solution pioneered by the Hathi Trust Research Center (https://www.hathitrust.org/htrc) is a set of non-consumptive research practices that strive to offer remote access to sensitive data without allowing users to abuse this access. One possibility is to allows users to run limited analyses and queries to extract features like sentiment scores from the data via an API or web interface, but only returns the sentiment scores or enough textual context to validate the scores without allowing access to the full text. Another possibility, called a Data Capsule (Zeng et al., 2014), allows a user to develop an analysis with limited access to the raw text, and then send the developed algorithm to a secure system where it can be run over the full corpus of data without giving the researcher any direct access to the data. In such a system, only limited and nonsensitive data can be returned to the researcher. Although these solutions can solve the problems of data access, they can be difficult to implement and cumbersome to use as they force the user to develop and validate analyses in an unfamiliar environment and possibly with different tools than they normally use.

Longer-Term Solutions Will Require Advocacy and Education

As surveyed above, research teams can already take a number of steps to mitigate the problems of data sharing between teams and within the community in general. However, none of these options is without problems. Better ways to share sensitive data will depend on concerted and long-term actions by the field as a whole. We think action should be taken in at least two directions.

Work Toward Better Data Agreements

The sharing of data between researchers does not pose a direct threat to the business models of news producers, and content archives like LexisNexis have a direct interest in ensuring that it remains possible to conduct and publish valid research with their materials. Thus, there might be scope to collaborate with these parties to work toward data agreements that allow for sharing raw data within open science principles. For this to happen, it is important that standard language be adopted to give these parties the confidence that we as scientists will not abuse these data or distribute them in ways that might hurt their profitability. This could include standard embargo periods (the value of news depreciates quickly) or a standardized procedure where local scientific archives such as DataVerse, the Inter-University Consortium for Political and Social Research (ICPSR), or the GESIS – Leibniz-Institute for the Social Sciences in Germany could give access to data on signing an appropriate agreement. The governments who ultimately fund most of our research could be convinced to pressure or regulate these data owners to allow scientific research by distributed teams, for example, as part of press subsidies or privileges or as a way to regulate the role of social media in the news ecosystem.

Work Toward Collaborative Open Data Sets

Progress in fields such as computational linguistics has profited tremendously from "shared tasks" where different research groups work on the same data set, such as the GigaWord dataset maintained by the Linguistic Data Consortium (https://catalog.ldc.upenn.edu/LDC2012T21). Similarly, in political communication shared data sets such as the Comparative Manifesto Project data have allowed researchers to collaborate and compare results. It would greatly help the research community studying political news to create and update open data sets of representative news articles. As the UK newspaper The Guardian has shown, open access does not necessarily mean a decline of income. It might be possible to use collective resources, either from funding agencies or the field itself, to convince a group of news publishers to open up part of their archives for research use after a reasonable time period has passed since publication. This could create a dataset on which different groups could develop, share, and validate their analyses and tools.

Conclusion

Given the complexities and uncertainties surrounding data sharing, there may be no single authority at your local institution who actually knows what the rules are, and many institutions lack a clear authority for making decisions about sharing of sensitive data. In our own experience, university lawyers tend to be very restrictive in their interpretations of the legal situation in order to minimize legal risks for the institution. Continued access to important library collections might be jeopardized, future ability for other researchers to use the same collections might be restricted, and lawsuits might be filed with substantial legal costs and large financial awards if cases go to trial. For these and other reasons, the easiest–and from a risk management standpoint, the best–decision is for the institution to simply say "No" and forbid the sharing of data altogether. A crucial factor in the success of cross-national collaborations using sensitive data might therefore be differences in "risk culture" among the collaborating institutions and their willingness to support researchers where legal boundaries are unclear and constantly evolving. From our experience, therefore, it would be beneficial for universities and research institutions to create and empower "data ombudspersons" to whom researchers could turn in cases of doubt, and – crucially – who could talk to each other directly across institutional boundaries. Data ombudspersons could shift the frame from risk prevention to research promotion and would relieve researchers from acquiring half-baked legal knowledge themselves.

As a research community, we need to find a sustainable and ethical solution for the problem of sharing our privates. We can't get away with ignoring the problems–as some currently seem to prefer, who are risking the entire research community's long-term access opportunities for individual short-term publication gain–but we also can't just give up open science. This will require increased awareness for the need to ethically share sensitive data among researchers, but also a concerted effort by the field to develop the relevant practices and standards required to do so, and to convince funding agencies, data owners, and regulators of the need to change agreements and regulations in ways that allow for open science practices to flourish.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft [WE 2888/7-1]; National Endowment for the Humanities [HJ-253500-17]; Nederlandse Organisatie voor Wetenschappelijk Onderzoek [463-17-004].

Notes on contributors

Wouter van Atteveldt is Associate Professor in Political Communication at the Vrije Universiteit Amsterdam, The Netherlands. His research interests include computational methods, automatic text analysis, digital tracking and the production, consumption and effects of political news.

Scott Althaus is Director of the Cline Center for Advanced Social Research, Merriam Professor of Political Science, and Professor of Communication at the University of Illinois Urbana-Champaign. His research interests explore the communication processes that support political accountability in democratic societies and that empower political discontent in non-democratic societies.

Hartmut Wessler is Professor of Media and Communication Studies at the University of Mannheim. His research interests include comparative political communication in online and offline media and normative assessment of democratic media performance.

References

Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Outlining the way ahead in computational communication science: An introduction to the ijoc special section on "computational methods for communication science: Toward a strategic roadmap". *International Journal of Communication*, *13*(9), 3876–3884. https://ijoc.org/index.php/ijoc/article/view/10533.

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., … Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology*, *4*(1), 1–15. https://doi.org/10.1525/collabra.158

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Karlan, D., & Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. https://doi.org/10.1177/0002716215572084

Van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2–3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Toward open computational communication science: A practical road map for reusable data and code. *International Journal of Communication*, *13*(20), 3935–3954. https://ijoc.org/index.php/ijoc/article/view/10631

Zeng, J., Ruan, G., Crowell, A., Prakash, A., & Plale, B. (2014, June). Cloud computing data capsules for non-consumptiveuse of texts. In *Proceedings of the 5th ACM workshop on Scientific cloud computing* (pp. 9–16). https://doi.org/10.1145/2608029.2608031