CESSDA Work Plan 2020

New Data Types

D2b: Report on the webinar "Archiving Social Media Data – Challenges and Proposed Solutions"

Document info

Dissemination Level	PU
Due Date of Deliverable	31/07/20
Actual Submission Date	16/07/20
Type	Report
Approval Status	Approved by CESSDA Tools & Services Working Group leader Mari Kleemola and CESSDA Training Working Group leader Irena Vipavc Brvar
Version	V1.3
Number of Pages	13
DOI	10.5281/zenodo.4672172

Version history

Version	Date	Comment	Revised by
0.1	12/07/20	Document created + first content	Kerrin Borschewski
0.2	13/07/20	First draft completed	Johannes Breuer
0.3	14/07/20	Language edits & formatting	Kerrin Borschewski
0.4	14/07/20	Internal peer-review by Libby Bishop (GESIS)	Johannes Breuer / Kerrin Borschewski
1.0	16/07/20	Final draft submission to CESSDA MO	Johannes Breuer / Kerrin Borschewski
1.1	27/08/2020	Addressed review comments by CESSDA MO	Johannes Breuer
1.2	16/10/2020	Addressed comments by CESSDA MO/WGL	Kerrin Borschewski
1.3	25/11/2020	Addressed further review comments from CESSDA MO/WGL + final editing & formatting	Johannes Breuer

Author List

Organisation	Name	Contact information
GESIS – Leibniz Institute for the Social Sciences	Johannes Breuer	johannes.breuer@gesis.org
GESIS – Leibniz Institute for the Social Sciences	Kerrin Borschewski	kerrin.borschewski@gesis.org

Peer-review

Organisation	Name	Contact information
GESIS - Leibniz Institute for the Social Sciences	Libby Bishop	ElilzabethLea.Bishop@gesis.org

Contents

List of Tables

Executive Summary

This report summarizes the relevant information on the CESSDA New Data Types webinar as an additional written record to the video recording and the published slides. The webinar "Archiving Social Media Data – Challenges and Proposed Solutions" was conducted on the 4th of June 2020. This webinar is part of the CESSDA New Data Types Workplan for 2020.

The first section of the report contains overview information on the webinar. The second part provides details on the content that was presented during the webinar. The third part captures questions from attendees as well as the responses to those questions. Section four presents a conclusion in which the main outcomes of the webinar and its implications are discussed, and section five comprises suggestions for further reading.

Abbreviations and Acronyms

ACLU	American Civil Liberties Union
ADP	Slovenian Social Science Data Archive
API	Application Programming Interface
CESSDA MO	CESSDA Main Office
DDI	Data Documentation Initiative
DPC	Digital Preservation Coalition
FAIR	Findable Accessible Interoperable Reusable
GESIS	GESIS – Leibniz Institute for the Social Sciences
ICPSR	Inter-university Consortium for Political and Social Research
ID	Identity
IPR	Intellectual Property Rights
QDR	Qualitative Data Repository at Syracuse University
RDA	Research Data Alliance
RDF	Resource Description Framework
SERISS	Synergies for Europe's Research Infrastructures in the Social Sciences
SOMAR	Social Media Archive
SP	Service provider
ToS	Terms of Service
U.S.	United States
URL	Uniform Resource Locator

1) General information on the webinar

On Thursday the 4th of June 2020 (3 pm - 5 pm CEST), the CESSDA New Data Types project conducted a webinar titled "Archiving Social Media Data – Challenges and Proposed Solutions" as part of its work plan for 2020. The webinar was organized by Johannes Breuer and Kerrin Borschewski (both GESIS). Delivery partner ADP was responsible for the registration and evaluation of the webinar and was also in charge of the technical coordination of the webinar (rehearsal, assigning speaker roles, monitoring the question chat, etc.). The webinar was delivered via GoToWebinar. Presentation of the webinar content was originally planned by GESIS (Johannes Breuer & Kerrin Borschewski) as a panel for the IASSIST 2020 conference (for which it was accepted). However, due to COVID-19, the IASSIST conference was moved to 2021. Hence, the decision was made to present the content in a form of a webinar. Considering the experience of the project team in webinar production, the implementation of this change of format went smoothly. All speakers who were part of the accepted IASSIST panel also agreed to participate in the webinar.

The speakers in the webinar were:

- Johannes Breuer (GESIS - Leibniz Institute for the Social Sciences)
- Kerrin Borschewski (GESIS - Leibniz Institute for the Social Sciences)
- Libby Hemphill (Inter-university Consortium for Political and Social Research (ICPSR))
- Janez Štebe (Slovenian Social Science Data Archive (ADP))
- Sara Day Thomson (University of Edinburgh)
- Elizabeth Lea (Libby) Bishop (GESIS - Leibniz Institute for the Social Sciences)
- Sebastian Karcher (Qualitative Data Repository (QDR) at Syracuse University)
- Oliver Watteler (GESIS - Leibniz Institute for the Social Sciences)

a) Information on participants

A total of 206 people registered for the webinar. The countries where the registrants came from are: Australia, Austria, Belgium, Canada, China, Czech Republic, Denmark, Finland, France, Germany, Greece, India, Ireland, Italy, Japan, Netherlands, New Zealand, Nigeria, North Macedonia, Norway, Peru, Philippines, Poland, Portugal, Senegal, Slovenia, Sweden, Switzerland, Taiwan, Thailand, United Kingdom, United States. A total of 142 registrants attended the live webinar. Table 1 presents an exact breakdown of where the registrants came from.

Table 1: Registration data per country

Country	**Number of registrants**
Unknown	3
Australia	1
Austria	3
Belgium	1

Canada	15
China	1
Czech Republic	1
Denmark	1
Finland	2
France	2
Germany	15
Greece	5
India	1
Ireland	7
Italy	3
Japan	1
Netherlands	4
New Zealand	1
Nigeria	3
North Macedonia	1
Norway	6
Peru	1
Philippines	1
Poland	1
Portugal	1
Senegal	1
Slovenia	6
Sweden	5
Switzerland	4
Taiwan	1
Thailand	1
United Kingdom	10
United States	97
Total	206

b) Materials on the webinar

Materials on the webinar are available on:

- Slides on Zenodo: http://doi.org/10.5281/zenodo.3875963
- Video on Zenodo: http://doi.org/10.5281/zenodo.3875963
- Video on CESSDA Training YouTube channel: https://youtu.be/EPP153H2Jow

2) Webinar content

The webinar consisted of two parts:

1. In the first part of the webinar, a group of invited experts gave short presentations of their current work related to the archiving of social media data.
2. The second part of the webinar was a moderated roundtable discussion during which the experts responded to questions from the attendees.

At the beginning of the webinar, the organizers Johannes Breuer and Kerrin Borschewski (both from GESIS) briefly introduced the topic and the speakers. This introduction was followed by 10 minute presentations, given by the invited experts. The first presentation was by Libby Hemphill from ICPSR who presented on the efforts at ICPSR to archive social media data. The following presentation was delivered by Janez Štebe (ADP) on "The application of FAIR Data Maturity Model to social media archiving." Sara D. Thomson (University of Edinburgh) then presented on "Shared strategies for ethical collection building" (for social media data). The fourth expert presentation was by Libby Bishop (GESIS) on "Archiving Social Media - Ethical challenges for data repositories". Sebastian Karcher (QDR) presented on "Archiving Qualitative Social Media Data." Finally, Oliver Watteler discussed legal issues related to archiving social media data in his contribution. The following subsections summarize the content of each presentation.

The presentations were followed by a roundtable discussion in which the panelists engaged with questions from the attendees.

a) Libby Hemphill (ICPSR) – Social media data archiving at ICPSR

While social media data is similar to other types of data in many regards – for example, in the sense that processing and documenting these data requires a substantial amount of effort – it also differs from the data that social scientists and social science data archives work with in several important ways. These relate to the properties of social media data (scale, speed of emergence/development, structure), the practices of handling it (required documentation, storage solutions), and ethics (privacy, privately owned data, etc.). ICPSR already holds quite a few social media data collections and has been developing its own Social Media Archive (SOMAR) infrastructure. An important question regarding the archiving of social media data is what the metadata should look like and how much observation-level indexing is necessary for the data to be (re-)useable. Other critical challenges for SOMAR at ICPSR relate to dealing with the Terms of Service (ToS) of social media platforms, and the need for technical and computational resources for storing social media data and making them accessible.

b) Janez Štebe (ADP) – The application of the FAIR Data Maturity Model to social media archiving

The FAIR Data Maturity Model is an operationalization of FAIR data principles by defining the fine-grained attributes of metadata and data to qualify as Findable, Accessible, Interoperable, and Reusable. This model has been applied to evaluate exemplary cases of archived social media data. The examples were selected based on a literature review and chosen to cover different types of social media data as well as different repositories. Out of the list of suggested cases, four have been assessed so far. There were substantial differences in the coverage of the FAIR data principles across the datasets. All four cases scored quite high on the Accessibility dimension (on average, 94% of the criteria for this dimension were met), and the overall scores for Reusability (80%) and Findability (82%) also were rather high. However, the average score for Interoperability was much lower (42%). In the future, more existing datasets will be assessed using this model, and the model may have to be adapted (e.g., by identifying the most relevant indicators per dimension or assigning weights to them).

c) Sara D. Thompson (Edinburgh University) – Shared strategies for ethical collection building

The *Web Archiving & Preservation Working Group* within the Digital Preservation Coalition (DPC) provides a forum for sharing experiences, establishing common goals, and informing policy development. With regard to social media data, this group arrived at several shared principles. First of all, social media data constitutes a valuable and critical asset. Secondly, while social media platforms do not have a mandate or obligation to preserve these data, collecting institutions do. Ethical decisions related to archiving social media data are not one size fits all. However, a shared ethical framework can support more confident collecting. One shared strategy for ethical collection building is a contextual ethical review that considers relevant aspects, such as the purpose of collecting, user awareness and consent, the legal and regulatory environment, but also the ethical mandate to collect. Notably, it is unethical, in some contexts, not to archive social media data.

d) Libby Bishop (GESIS) – Ethical challenges for data repositories

The ethical challenges for social media data can be straightforward if they are public, and the information they contain is minimal. An example of this would be the names, user names/IDs, and party names of politicians who are active on Twitter in the course of a specific campaign[1]. However, if the data have disclosure risks, implementing access controls

[1] Political Campaigning on Twitter During the 2019 European Parliament Election Campaign, https://doi.org/10.7802/1.1995 (date of access: 27/08/2020)

can be one solution[2]. In general, social media data should be archived as archives claim broad social responsibilities, e.g., related to preservation (data can have historical value) and reproducibility of research outputs (ensuring the integrity of data and methods). Complying with ToS is generally an important goal for archives. However, archives also have duties and obligations, some of which may compete with complying with ToS (e.g., documenting historically relevant events or movements). Also, in many cases, ToS can be interpreted differently, and, in practice, institutions sometimes treat them differently. While there are no universal ethical guidelines for archiving social media data, there are some useful resources that provide guidance (e.g., in the paper by Williams et al., 2017[3]: or in Appendix A of the SERISS WP6-D3 Report[4]).

e) Sebastian Karcher (QDR) – Archiving Qualitative Social Media Data

Qualitative social media data have some distinctive characteristics. They are typically manually collected (not using an API), have a small sample size, and may include a range of different sources within a single project. One challenge for archiving this kind of data is that (parts of) it may be deleted from the original sources (social media platforms). To ensure that the data can still be accessed tools like *perma.cc*[5] or the *Internet Archive*[6] can be used. QDR has developed the package *archivr*[7] for the statistical programming language R that can be used to extract URLs and archive them using the Internet Archive or perma.cc. A toll that QDR uses to make videos and other non-static web content formats available is *webrecorder.io*.[8] Overall, the effort of sharing qualitative social media data faces similar challenges to sharing larger-scale social media data. However, given the smaller number of items, individual curation & checks are more feasible. Importantly, some of the tools that can be used for archiving qualitative social media data may not scale well. In such cases, the tools offered by Documenting the Now[9] may be useful.

---

[2] Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness, https://doi.org/10.7802/1166 (date of access: 27/08/2020)
[3] Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140
[4] SERISS Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data, https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies... FINAL.pdf (date of access: 27/08/2020).
[5] Perma.cc, https://perma.cc/ (date of access: 27/08/2020)
[6] The Internet Archive, https://archive.org/ (date of access: 27/08/2020)
[7] archivr R Package, https://github.com/QualitativeDataRepository/archivr/ (date of access: 27/08/2020)
[8] Webrecorder.io/Conifer, https://conifer.rhizome.org/ (date of access: 27/08/2020)
[9] Documenting the Now, https://www.docnow.io/ (date of access: 27/08/2020).

f) Oliver Watteler (GESIS) – Legal issues

Archiving consists of preservation, documentation, and publication of data. Before ingesting data, archives need to clarify the legal basis of a data collection and, thus, the rights to the data. They also need to clarify conditions for re-use. Those clarifications are necessary for an archive agreement. When they ingest the data, archives need to check them with regard to their quality as well as the content of both the data and their documentation, while also keeping in mind legal issues. Based on the outcomes of these checks, information may have to be reduced, or access may have to be restricted. There are typically different rights involved when working with social media data: Contractual agreements between user and platform provider (ToS), data protection for personal information, intellectual property rights for content like photos, videos, audio, (creative) text, and database rights. Importantly, there are many different types of social media data, and they can be collected in various ways. Data can, e.g., be purchased from platforms of third-parties, collected via APIs or through web scraping. All of these methods are associated with different legal regulations. There may also be different legal bases for the collection of social media data. Apart from the general freedom of research, which differs between countries, another common basis in social science research is informed consent. Another relevant basis for social media data is the agreement to ToS (e.g., through a usage agreement or purchasing contract). While in the ideal case, all rights related to the data are clarified before they are archived, this is often not the case for social media data. For archives, this means that they need to develop new procedures for dealing with this.

3) Questions and Answers from Q&A session

Published as a separate document.

4) Conclusion

The webinar had several aims. First of all, the "New Data Types" project wanted to reach out to CESSDA SPs and others interested in the topic of archiving social media data. Furthermore, the webinar was supposed to provide means for a discussion with the community and to provide or develop answers for open questions regarding the archiving of social media data. The Q&A session proved to be especially valuable for this. The third aim of the webinar was to establish a network of archivists and researchers who are active in the area of archiving social media data. Such networks help CESSDA to stay informed about the experiences, challenges, and ongoing work on archiving social media data. The participation of people from institutions that are not CESSDA members (ICPSR & QDR) also broadened the perspective and provided insights on how the challenges in archiving social media data are addressed at other institutions and what procedures and solutions they have been developing.

Overall, the aims of the webinar were reached. The very high number of registrants is proof of the broad interest in and the topic. Given that the participants came from all over the world, the webinar was also a suitable platform for promoting the ongoing work within CESSDA on new data types. For future CESSDA work on new data types, we see two promising avenues based on the experiences from the webinar: 1) To more systematically assess the experiences and needs of researchers working with social media data (and maybe also other kinds of new data, such as digital trace data more broadly), it may be helpful to conduct a survey among this target group. 2) In addition to the output from the 2020 project on "New Data Types", to provide further guidance, especially for archival staff working with new data types, it may be advisable to develop further resources that can be consulted by those who are new to the topic (e.g., online materials that can be updated and expanded).

5) Further reading

Bishop, L., & Gray, D. (2017). Chapter 7: Ethical challenges of publishing and sharing social media research data. In K. Woodfield (Eds.), *Advances in Research Ethics and Integrity* (pp. 159–187). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002007

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Hemphill, L., Hedstrom, M. L., & Leonard, S. H. (2020). Saving social media data: Understanding data management practices among social media researchers and their implications for archives. *Journal of the Association for Information Science and Technology*, 3, 34. https://doi.org/10.1002/asi.24368

Hemphill, L., Leonard, S. H., & Hedstrom, M. (2018). Developing a Social Media Archive at ICPSR. Proceedings of Web Archiving and Digital Libraries (WADL'18). http://hdl.handle.net/2027.42/143185

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), Advance online publication. https://doi.org/10.1177/2053951717736336

Mannheimer, S., & Hull, E. A. (2018). Sharing selves: Developing an ethical framework for curating social media data. *International Journal of Digital Curation*, *12*(2), 196–209. https://doi.org/10.2218/ijdc.v12i2.518

Thomson, S. D. (2016). *Preserving social media*. Digital Preservation Coalition. https://doi.org/10.7207/twr16-01

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and

algorithmic estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140