CESSDA Work Plan 2020

New Data Types

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

Questions and Answers from Q&A session

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	7
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

Questions and Answers from Q&A session 6

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 Work Plan 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) Questions and Answers from Q&A session

The attendees had several opportunities for posing questions. Before the webinar, it was possible to ask a question via the registration form by providing an answer to the question "Do you have any questions that you would like to discuss during the roundtable part of the webinar?". During the presentations and the roundtable discussion, attendees could pose their questions using the Chat/Questionbox provided by GoToWebinar. The panelists discussed the questions in the roundtable discussion. Transcriptions of the questions and (summarized) responses are presented in the following (with editing for language, brevity, and clarity):

1) Question for Libby Hemphill: Given your role with archiving ethnic minority data and the fact you are based in Michigan, what do you know of what is being collected about the current #Blacklivesmatter activity and what are the specific ethical or other challenges of that?
   Answer: The Documenting the Now group is very active in this area. A challenge in this area is something like solidarity through silence, such as the use of completely black profile pictures that replace regular profile pictures. ICPSR is not collecting data on this itself. ICPSR is currently discussing whether to build infrastructures for collecting data. Currently, the stance of ICPSR is that it should be preserving data that researchers think is important instead of deciding itself what data to collect.
2) Question for Libby Hemphill: In your presentation you talked about metadata enhancement. Could you maybe specify this a bit? (in our experience, we may have the feeling that metadata must be «perfect » when published)
   Answer: Metadata is, quite possibly, never perfect. Metadata can be updated, which may, e.g., be necessary to improve findability. One of the biggest challenges with regard to metadata for social media data is the documentation of provenance. It is essential to specify what information about how the data was collected is required. Importantly, archives may also add metadata that researchers cannot provide (for different reasons). For example, for the area of machine learning, the outputs of models may also be metadata that can be attached.
3) Question for Janez Štebe: Is there a minimum requirement for data to meet each of the FAIR principles? For example 80% or higher?
   Answer: There is no such requirement, but this is a topic of discussion, for example, in the Research Data Alliance (RDA) group that is active in this area. The standards or requirements might differ depending on the specific community that is addressed and the type of data. Related to that, it always depends on the specific case how much sense it makes to invest time and resources into, e.g., enhancing metadata or improving the documentation of the data in general. A key criterion here certainly is the (assumed) reuse value of the data.
4) Question: Is it legal to use images/videos shared on the social media platform? For example, profile pictures, screenshots showing tweet or Facebook posts

Answer: This question can relate to two things: 1) Images, videos, etc. shared on social media platforms and 2) screenshots of social media content. Intellectual property rights certainly play a role here. If, for example, somebody takes a photo and posts it on social media, that person is the author of that picture. This means that, legally, this person owns the rights to that picture. In order to reuse the picture it would be necessary to ask the author for permission. In practice, this is often not done but this is problematic. It is necessary to ask for permission to reuse unless there is a specific open license attached to it.

5) Question: What if researchers do not know which Terms of Service they agreed to? What if they scraped the data?
Answer: It is often possible to reconstruct which ToS were valid when the data were collected. In the case of scraping, it is quite likely that ToS are violated. In such cases, archives need to decide how to handle and what to do what this data.
What needs to be taken into account is the value of the data for science and history.
It is important to be aware of ToS and consider which parts of them are there to protect the interests of users. Importantly, archives have different mandates, obligations, and interests than companies that collect and use data for commercial services but also than individual researchers. There are also examples of research, mostly from the qualitative area, in which it is not possible to get consent from the people whose data were used. Hence, archives need to be able to deal with cases where specific guidelines are not (or cannot be) followed.
For the German legal system, the RatSWD has published a legal expertise[1] regarding the scientific use of web scraping. In several important regards, continental European law here is different from, for example, U.S. law. In European continental law, what is valid as ToS is typically regulated by national laws. There is also another legal expertise from Germany[2], and both conclude that, if certain criteria are met, web scraping for scientific purposes is legal and should be treated differently than web scraping for commercial purposes. A further key question in this context is who is or would be willing to take such cases to court. Individual researchers typically lack the resources to do so.
In the U.S., the American Civil Liberties Union (ACLU) wants to go to court, and they were recently successful. This shows that it is crucial to advocate for policy change concerning legal regulations and platform ToS. This is something that individual researchers alone can typically not do. Archives, however, are in a stronger position that allows them to do so.

[1] RatSWD [German Data Forum] (2020): Big data in social, behavioural, and economic sciences: Data access and research data management. *RatSWD Output 4* (6). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.52 (date of access: 27/08/2020).
[2] Golla, S. J., von Schönfeld, M. (2019). Kratzen und Schürfen im Datenmilieu – Web Scraping in sozialen Netzwerken zu wissenschaftlichen Forschungszwecken. *Kommunikation & Recht, 22*(1), 15-21. https://baecker.jura.uni-mainz.de/files/2019/01/KUR_01_19_Beitrag_Golla_Schoenfeld.pdf (date accessed: 27/08/2020).

6) Question: Where/how do copyright and 'legal restrictions' impact archiving social media data?
   Answer: Copyright varies across jurisdictions. In the U.S., and also many other countries, copyright protects original work. This protection typically does not apply to statements of fact. In U.S. law, there are conditions under which it is possible to claim "fair use" (e.g., if commercial interests of the original owners are not affected). For archives, this is a problematic area, as some things, e.g., related to "fair use" are poorly defined.
   Importantly, copyright regulations also vary across social media platforms.
7) Question: Who should archive social media data? For how long is the commitment? Where is the funding to curate the collections coming from?
   Answer: An essential question in this regard is whether there is a demand for this kind of data. Archives typically use their own (standard) funding as well as third-party funding to make social media data available. Funders usually want to know why archiving the data is relevant.
8) Question: Follow up question (to the question regarding the use of images/videos shared on social media): If, for example, we are working with 1000 Twitter profiles, it is not possible to ask everyone for permission to use their profile picture and name. In such a case, can we stop bothering about IPR and publish the results? Or what should we do?
   Answer: It is necessary to show/explain why this is not possible. If you can show that it was not possible (for you) to contact the people, it can be possible to archive the data.
   Another question is whether these (parts of) the data need to be archived. Often the profile pictures themselves are not part of the research question. Researchers should always ask themselves whether they need to archive the parts that are most creative (which makes them a potential copyright issue) or most identifying (which makes them a potential privacy risk).
9) Question: What are your experiences with social media researchers' interest in archiving their data? Are they, e.g., actively approaching archives? Especially authors of 'gold standard', high-quality datasets that Libby Bishop mentioned? From our experience some researchers from social science/media and communication backgrounds are, indeed, interested, but, e.g., the large georeferenced dataset that Libby (Bishop) mentioned needed to be very actively recruited - so considerable effort from various GESIS archivists was required.
   Answer: Depending, among other things, on the requirements of funders and journals, researchers in some disciplines and countries do not have to archive their data. Hence, many archives have to recruit and motivate researchers to archive their social media data. Many of the available archived social media collections are associated with specific publications. This suggests that requirements from journals play an important role.

At ICPSR, the experience regarding the interest of researchers in archiving their data is similar to that for other types of data. This means that, in many cases, the archive needs to reach out to them. For many, it is still not part of their standard research practices. However, there are also cases where researchers actively desire or want to archive their social media data. A concern that many researchers in this area have relates to ToS and how they are not sure whether they are allowed to archive the data or how they can be archived in line with the ToS of the platform for which the data was collected. Hence, many researchers are hesitant to archive their data.
Some researchers are also required to archive the data (e.g., by funders or journals) and need guidance on how to do this.

10) Question: What type of standards do you use to implement metadata? DDI, RDF, (Disco)
Answer: QDR, ICPSR, ADP, and GESIS use (various versions of) DDI. However, the question is if this is sufficient. For social media data, there may be a need to extend certain metadata fields.