

Archiving Social Media Data

Challenges and Proposed Solutions

Johannes Breuer; GESIS – Leibniz Institute for the Social Sciences

Kerrin Borschewski; GESIS – Leibniz Institute for the Social Sciences

June 4th 2020 / CESSDA webinar

DOI: 10.5281/zenodo.3875963

 [cessda.eu](https://www.cessda.eu)

 @CESSDA_Data



Licence: CC-BY 4.0

Agenda

- Introduction of Speakers
- Presentations Panelists
- Roundtable Discussion

The Organizers



Johannes Breuer

Johannes.Breuer@geis.org

@MattEagle09



Kerrin Borschewski

Kerrin.Borschewski@gesis.org

The Panelists



Libby Hemphill

libbyh@umich.edu

@libbyh



Janez Štebe

Janez.Stebe@fdv.uni-lj.si

@Janez_Stebe



Sara Day Thomson

sthoms13@exseed.ed.ac.uk

[@sdaythomson](https://twitter.com/sdaythomson)



Libby Bishop

ElizabethLea.Bishop@gesis.org

[@LibbyBishopPhi](#)



Sebastian Karcher

skarcher@maxwell.syr.edu

@adam42smith



Oliver Watteler

Oliver.Watteler@geis.org

ICPSR



CESSDA Webinar

Archiving Social Media Data: Challenges and Proposed Solutions

Libby Hemphill

Director, Resource Center for Minority Data

ICPSR



How is social media like other social science data?

1. Researchers worry about getting scooped
2. Preparing data for reuse takes a lot of effort
3. Found data requires special manipulation and documentation

What makes social media data special?

1. Data properties: scale, speed, structure
2. Data practices: finding, curating, sharing, and storing
3. Ethics: private owners, PII

Scale

ICPSR's existing archive: 8-9TB

1,979,707,993 Tweets

1403 months of Gab posts

15 months of Reddit comments

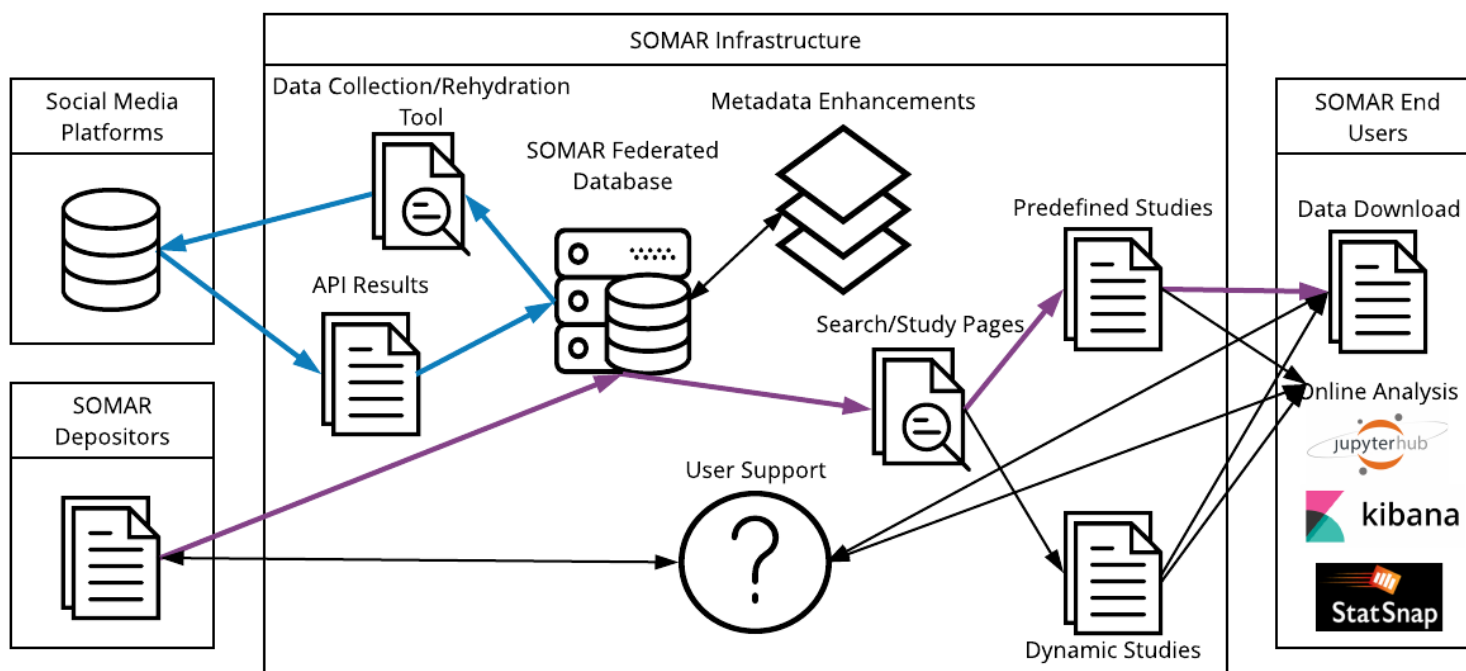
Structure

- What should metadata for social media data look like?
- How much should we invest in observation-level indexing?

Ethics

- When and how should we consider user's intent in making collection development decisions?
- Should SOMAR be all restricted access?
- How should we decide who can see the data and how it can be used?

Social Media Archive (SOMAR) at ICPSR



SOMAR's Main Challenges

- Metadata and structure
- Platforms' terms of service
- Technical and computational resources
- Privacy
- Costs/benefits

ICPSR Publications

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>

The application of FAIR Data Maturity Model to social media archiving

Overview of selected cases

Janez Štebe / CESSDA ADP, University of Ljubljana

June 4, 2020, 3 pm CEST / CESSDA webinar "Archiving Social Media Data – Challenges and Proposed Solutions"

 cessda.eu

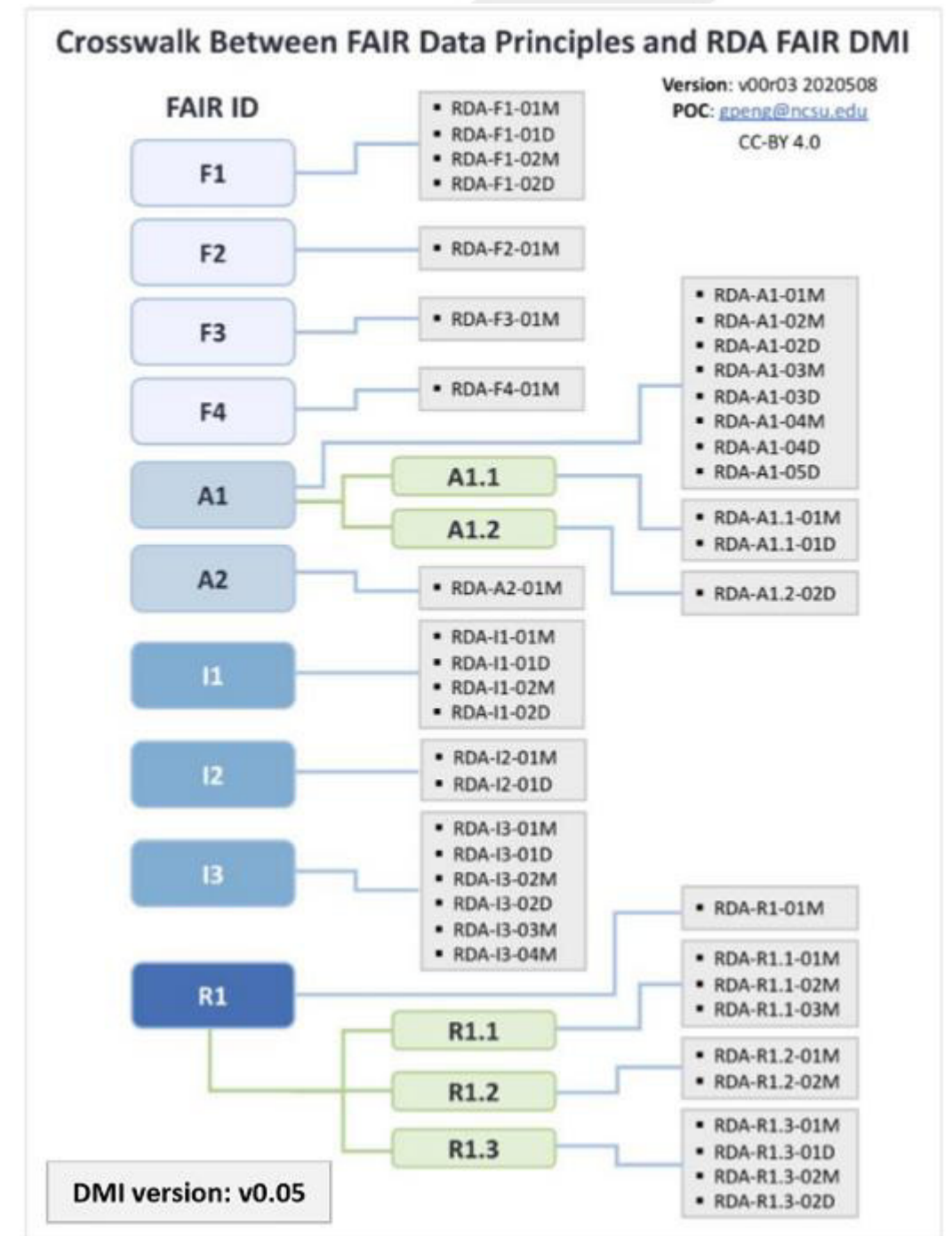
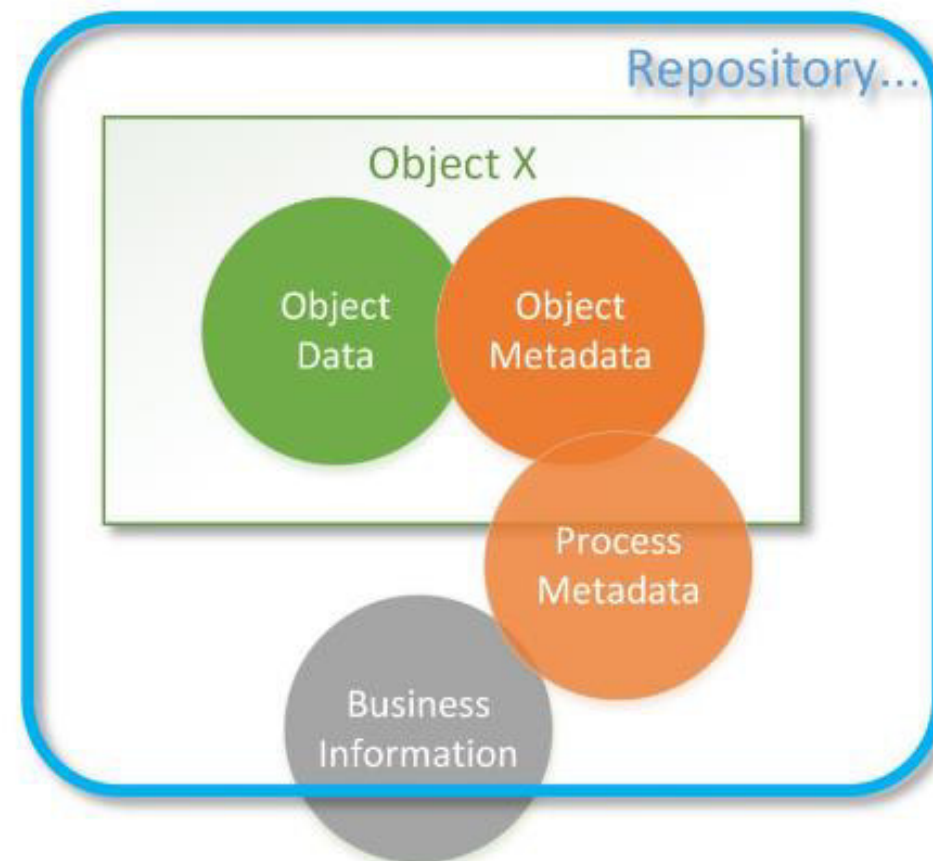
 @CESSDA_Data



Licence: CC-BY 4.0

FAIR Data Maturity Model

- ◊ Operationalisation of FAIR Data Principles by defining the fine grade attributes of **metadata** and **data** to qualify as Findable, Accessible, Interoperable and Reusable
- ◊ FAIR to be considered through data lifecycle stages
 - ◊ data repository characteristics are important
 - ◊ CTS + FAIR (community specific)



Ge Peng (2020): Evaluating the FAIRness of Environmental Data. Application of the RDA FAIR Data Maturity Indicators. #9 Workshop of the RDA FAIR Data Maturity Model Working Group, May 20–21, 2020 . https://www.rd-alliance.org/system/files/documents/20200520_FAIR_WG_slides_v0.04.pdf

Hervé L'Hours, Anusuriya Devaraju, Ilona von Stein, & Mustapha Mokrane. (2020, May 14). FAIRsFAIR Comments Response on RDA FAIR Data Maturity Model Working Group (2020). FAIR Data Maturity Model: specification and guidelines (Version 01.00). Zenodo. <http://doi.org/10.5281/zenodo.3827109>

SERISS project WP6: New forms of data – legal, ethical and quality issues

- ◇ ADP: Linked third party CESSDA partner
- ◇ Reports and Guidelines: <https://seriss.eu/resources/deliverables/>
- ◇ Overview around **ethical issues**
 - Consent (when it's necessary, if practical, and how to obtain)
 - Public or private communication
 - Confidentiality: Anonymization, sensitive content, vulnerable population
- ◇ **Legal issues**
 - Terms of Services (changing, limiting use, when to and when not to follow)
 - Licences and permissions of use (Controlled access to sensitive data, evaluating purpose of use to balance the risks)
 - GDPR (rules and exceptions for research)



seriss

SYNERGIES FOR EUROPE'S
RESEARCH INFRASTRUCTURE
IN THE SOCIAL SCIENCES

Deliverable Number: 6.3

Deliverable Title: Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data

Work Package: 6 New forms of data: legal, ethical and quality matters

D6.3-Report-on-legal-and-ethical-framework-and-strat... 3 / 47

- 3.1.8 Intellectual property
- 4 Publishing and sharing data collected from social media**
 - 4.1 Ethical issues
 - 4.1.1 Informed consent and unconsented data
 - 4.1.2 Shift from anonymisation to risk minimisation
 - 4.1.3 Public or private or...
 - 4.1.5 Research integrity
 - 4.2 Legal issues
 - 4.2.1 IPR protection
 - 4.2.2 Privacy considerations
 - 4.2.3 Social media platforms' terms and conditions

2

4.3 Sharing of social media data

4.4 Survey on social media data in European data archives

Definitions

References

Appendix A: Short guide on legal and ethical issues for the when using social media for research

CESSDA Work Plan Tasks Project: New Data Types 2020

- ◊ Leading Partner GESIS
- ◊ Challenge:

„Led by the FAIR principles, these new data need to be made findable, accessible, identifiable, and re-usable by adapting and extending infrastructures designed for traditional data sources.“

- ◊ WP Project starts from where some previous projects ended:
 - ◊ Assessment of examples of sharing SM data

Preliminary work around FAIR & SM data

Aim:

- ◊ Pilot test **FAIR Data Maturity Model (DMM)**- produced by RDA WG, 2019-2020:
<https://www.rd-alliance.org/groups/fair-data-maturity-model-wg>
- ◊ To demonstrate the usefulness of evaluating FAIR DMM on selection of SM data sharing cases

Expected outcome:

Orientation for future work

- about the need to adapt FAIR DMM for the purpose of SM evaluation, and
- about potential results of assessment of practice of SM data archiving

Selection of SM data archiving cases

- ◊ Based on literature overview (CESSDA DMEG, SERISS report, Kinder-Kurlanda et al., Mannheimer&Hull, DocNow)
- ◊ Covering following categories of cases:
 - ◊ Typical, exceptional, problematic
- ◊ and Repository services:
 - ◊ Thematic social media data sources, General repositories and Disciplinary data centres

List of suggested cases / 4 preliminary assessed

DO_ID	Link	Repository
TAttacks	https://doi.org/10.18712/nsd-nsd2434-v1	CESSDA/NSD
UK_2015_G_E	https://dx.doi.org/10.5255/UKDA-SN-852772	CESSDA/reshare.ukdataservice
German_Bundestag_E_2013	https://doi.org/10.4232/1.12319	CESSDA/GESIS Data Archive
Geotagged_Twitter	http://doi.org/10.7802/1166	CESSDA/data.gesis.org
Janes_Tweet	http://hdl.handle.net/11356/1142	CLARIN.SI/repository
TweetsKB	http://doi.org/10.5281/zenodo.1629949	ZENODO
News_Sharing	https://doi.org/10.7910/DVN/5XRZLH	Harvard Dataverse
WOI_2018	https://doi.org/10.7910/DVN/YMJJFC	Harvard Dataverse
T3	https://dataverse.harvard.edu/dataverse/t3	Harvard Dataverse
OkCupid	https://osf.io/5qwr8/	OSF
Occupy	https://doi.org/10.5061/dryad.q1h04	DRYAD
In_the_mood	https://doi.org/10.5061/dryad.5302r	DRYAD
COVID_19	http://dx.doi.org/10.21227/781w-ef42	ieee-dataport
#metoo_project	https://www.schlesinger-metoo-project-radcliffe.org/access-the-collection	Thematic social media data sources
GeoCoV19	https://crisisnlp.qcri.org/covid19	Thematic social media data sources

Summary results of FAIR DMM assessment

Counts of 1 – present / 0 – absent for each of the FAIR DMM indices divided by N indices

DO_ID Case	F	A	I	R	Fair Mean over groups	Repository
In_the_mood	57%	92%	33%	90%	68%	DRYAD
News_Sharing	100%	92%	17%	90%	75%	Harvard Dataverse
TAttacks	100%	100%	33%	70%	76%	CESSDA/NSD
UK_2015_G_E	71%	92%	83%	70%	79%	CESSDA/reshare.ukdataservice
Total Mean over cases	82%	94%	42%	80%	74%	
N indices per group	7	12	12	10	41	

**RDA-R1.1-02M :
Standard reuse
licences**

**RDA-R1.1-03M :
Metadata refers to a
machine-
understandable
reuse licence**

Partially absent from F: F1-01D Data is identified by a persistent identifier

Absent from I: 1-02D Data uses machine-understandable knowledge representation

Illustration of one of the Accessible points

RDA-A2-01M Metadata is guaranteed to remain available after data is no longer available

UK_2015_G_E (Reshare UK DS):

The Data Service Provider shall: Retain the right to remove all or any part of the Data Collection if it is found to be in breach of the law. A metadata record that cites the Data Collection will remain visible.

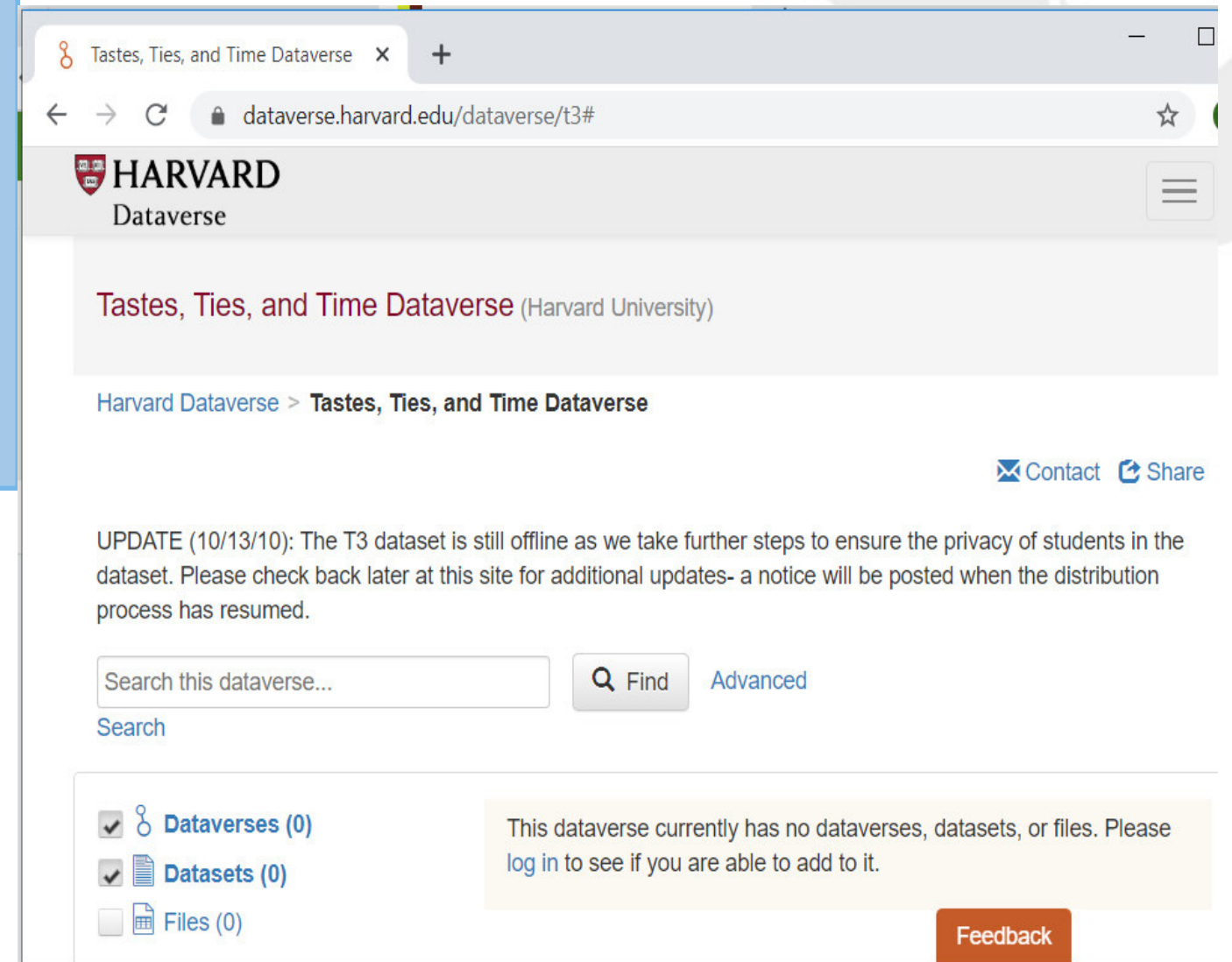
<http://reshare.ukdataservice.ac.uk/legal/#Deposit> ReShare Legal Documentation



The screenshot shows the NSD website interface. The main content area displays the following metadata for the dataset 'Use of Social Media During and After the Terrorist Attacks in Norway in 2011, 2017':

- Time Method:** Cross-sectional survey
- Data Collector:** Hornmoen, Harald, Oslo and Akershus University College of Applied Sciences (HIOA)
- Sampling Procedure:** Twitter messages from Norway in July and August 2011 were collected from Gnip, which manages Twitter's historical archive. In total, the data consists of 2.2 million Twitter messages posted between July 20th and August 28th, 2011 and represents all of the Norwegian Twitter sphere that was possible to collect during this period. The search was based on a Boolean search where language recognition, location information and similar parameters, were the conditions. The Twitter message is inserted into a SQL database for analysis. Facebook feeds from survivors on Utøya from the same period were collected with consent. The interview data consists of semi-structured interviews with survivors from Utøya, information and communication workers at institutions such as PST, the Directorate of Health, the Public Health Institute, the Directorate for Social Security and Emergency Planning, the Police Directorate, Oslo Police District, Oslo University Hospital, and the National Center for Health Services' Communication Preparedness. Text data from online newspapers and newspapers were collected via Retriever based on searches for articles published July 22 and July 23, 2011, that refer to social media. Some online news data was also collected from the Danish media researcher Aske Kammer, who downloaded the online newspaper's front pages every 5 minutes on the 22nd of July 2011.
- Mode of Data Collection:** (partially visible)

Hornmoen, H. (2017). Use of Social Media During and After the Terrorist Attacks in Norway in 2011, 2017 [Data set]. NSD – Norwegian Centre for Research Data.



The screenshot shows the Harvard Dataverse website for the 'Tastes, Ties, and Time Dataverse'. The page includes the following information:

- Harvard Dataverse logo and navigation menu.
- Dataset title: Tastes, Ties, and Time Dataverse (Harvard University)
- Breadcrumbs: Harvard Dataverse > Tastes, Ties, and Time Dataverse
- Contact and Share buttons.
- Update notice: UPDATE (10/13/10): The T3 dataset is still offline as we take further steps to ensure the privacy of students in the dataset. Please check back later at this site for additional updates- a notice will be posted when the distribution process has resumed.
- Search bar: Search this dataverse... Find Advanced
- Summary statistics: 0 Dataverses, 0 Datasets, 0 Files.
- Message: This dataverse currently has no dataverses, datasets, or files. Please log in to see if you are able to add to it.
- Feedback button.

Tastes, Ties, and Time Dataverse:
<https://dataverse.harvard.edu/dataverse/t3>

Future work

Possible improvements of measurement tool:

- ◊ Selection of most relevant indicators or assign weights
- ◊ Indicators that are hard to assess:
 - ◊ Need to establish specific operationalisation of maturity levels for certain indicators depending on the SM data types characteristics:
 - ◊ E.g. Reusable **R1.2 RDA-R1.2-02M** Metadata includes provenance information according to a cross-community language ● Useful

0 – not applicable	
1 – not being considered this yet	
2 – under consideration or in planning phase	
3 – in implementation phase	
4 – fully implemented	Need to agree upon what is the ultimate goal?

Aspirational FAIR sharing framework

- ◇ *Original FAIR Guiding Principles (<https://doi.org/10.1038/sdata.2016.18>):*
 - ◇ *,The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings.'*



Thank you for attention!

Questions?

Janez.Stebe@FDV.Uni-Lj.Si

 cessda.eu

 @CESSDA_Data



Licence: CC-BY 4.0



THE UNIVERSITY *of* EDINBURGH



CESSDA WEBINAR
*ARCHIVING SOCIAL MEDIA DATA –
CHALLENGES AND PROPOSED SOLUTIONS*
4 JUNE 2020



SHARED STRATEGIES FOR
ETHICAL COLLECTION BUILDING



Sara Day Thomson
Digital Archivist
Previously: Research Officer, DPC

Overview

- Background to the DPC *Web Archiving & Preservation Working Group*
- Different Missions & Mandates within WAPWG
- Applying Ethical Review in Different Contexts
- Shared Strategies for Ethical Collection Building



Digital Preservation Coalition's *Web Archiving & Preservation Working Group*



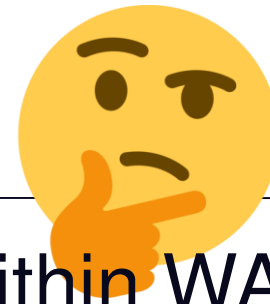
Forum for participants to...

- Share experiences
- Establish common goals
- Inform their own policy development
- Engage with a mutually supportive environment



THE UNIVERSITY of EDINBURGH





Different Missions & Mandates within WAPWG

Researcher Generated Datasets

- Collected for a particular research question
- Machine-readable datasets for quantitative analysis
- Possible platform 'archives', web captures, or even screenshots

Heritage Collections

- Based on collecting policies
- Support range of research needs
- Record of major events, groups, or themes (e.g. COVID-19 collecting)
- Gallery exhibitions or artworks, social media as medium

Legal or Regulatory Compliance

- Government or Business archive
- Evidential value
- Demonstrate transparency and accountability
- Marketing or promotional content
- Metrics or consumer analysis



Applying Ethical Review in Different Contexts

Web Archiving & Preservation Working Group: Social Media & Complex Content

16 January 2020

National Records of Scotland, Edinburgh



Small Group Breakout Activity

Ethical Deliberation:

To Archive Twitter or Not to Archive Twitter

(Observed) Shared Principles

- Social media data constitutes a valuable and critical asset
- Platforms do not have a mandate or obligation to preserve these data, but collecting institutions do
- Requirement to archive social media supersedes the challenges of collecting it
- Ethical decisions are not one size fits all, BUT
- Shared ethical framework will support more confident collecting



Shared Strategies for Ethical Collection Building

Contextual ethical review that considers:

1. Purpose of Collecting
2. User Awareness and Consent
3. Subjectivity and (Un)conscious Bias
4. Legal and Regulatory Environment
5. Ethical Mandate to Collect
6. Platform and Functionality
7. Access and Use/Re-use



It is unethical, in some contexts, not to archive social media data.





DocNow

Recommendations for Ethical Collecting

- (Educate social media users and researchers who might become depositors about terms of service)
- Engage and work with relevant communities
- Generate documentation beyond what can be collected without permission online
- Comply with platforms' terms of service where they are congruent with the values of the communities being documented
- When possible, apply traditional archival practices such as appraisal, collection development, and donor relations



Good practice is in the process





THE UNIVERSITY *of* EDINBURGH



SARA DAY THOMSON
DIGITAL ARCHIVIST
STHOMS13@ED.AC.UK
@SDAYTHOMSON

Influencing the world since 1583

Archiving Social Media - Twitter

Ethical challenges for *data repositories*

Libby Bishop *GESIS–Leibniz-Institute for the Social Sciences*

4 June 2020 *Archiving Social Media Data – Challenges and Proposed Solutions* *CESSDA Webinar*

 cessda.eu

 @CESSDA_Data



Licence: CC-BY 4.0

Easy – IF info minimal, or public

```
_id,userid
366711188735799296,1117526742
366711604059963394,402220351
366712022802509828,378242224
366712171029204993,23762413
366712387207831553,56767791
366712390856880129,378242224
366712516950245377,56767791
366712647019794432,56767791
366712800418086912,23762413
366712903572791298,24887580
366712909251883008,402220351
366712931599138817,56767791
366712938930778112,378242224
366714297038028801,378242224
366715949375696896,267772916
366716324908511232,378242224
366716633047236609,378242224
366716773317349379,267772916
366716796755116032,1189268659
366719155732348928,123549610
366719278008905729,123549610
366720681876008961,90046967
366721461752311810,421128471
366722293239517184,49976840
366722870145073152,455624784
```

[Home](#) / [Data catalogue](#) / [Studies](#) / [Study](#)

Tweets used to study reports of food fraud related to fish products 2018

Details

Access data

Details

Title:	Tweets used to study reports of food fraud related to fish products 2018
Study number (SN):	853378
Access:	These data are open
Persistent identifier:	10.5255/UKDA-SN-853378
Principal investigator(s):	Edwards, P, University of Aberdeen Markovic, M, University of Aberdeen Petrunova, N, University of Aberdeen Chenghua, L, University of Aberdeen Corsar, D, University of Aberdeen

Sponsors and contributors

Political Campaigning on Twitter During the 2019 European Parliament Election Campaign

Cite as

URI	https://doi.org/10.7802/1.1995
Primary Researcher:	Stier, Sebastian; GESIS - Leibniz-Institut für Sozialwissenschaften Popa, Sebastian A.; Newcastle University Braun, Daniela; LMU Munich
Publisher:	GESIS - Leibniz-Institute for the Social Sciences
Publication Year:	2020
Availability:	Free access (without registration)
Project funder:	VolkswagenStiftung
Replication Server:	No

Content

Subject Area: European Politics
Political Process, Elections, Political Sociology, Political Culture

Font	Alignment	Number	Styles	Cells	E
country;name;party;english_party_name;gender;incumbent;place_list;twitter_screenname;twitter_id;parlgov_id;ches					
B	C	D	E	F	G
;party;english_party_name;gender;incumbent;place_list;twitter_screenname;twitter_id;parlgov_id;ches_id;ees_party_id;eu_actor;eu_pa					
rald;FPÄ–;Austrian Freedom Party;Male;0; 1;vilimsky;303234771; 50;1303;1040420;EP Candidate;MENF					
. Georg;FPÄ–;Austrian Freedom Party;Male;1; 2;georgmayermep;2821282972; 50;1303;1040420;EP Candidate;MENF					
tra;FPÄ–;Austrian Freedom Party;Female;1; 3;NA;NA; 50;1303;1040420;EP Candidate;MENF					
man;FPÄ–;Austrian Freedom Party;Male;0; 4;NA;NA; 50;1303;1040420;EP Candidate;MENF					
sna;FPÄ–;Austrian Freedom Party;Female;0; 5;NA;NA; 50;1303;1040420;EP Candidate;MENF					
sabeth;FPÄ–;Austrian Freedom Party;Female;0; 6;NA;NA; 50;1303;1040420;EP Candidate;MENF					
sef;FPÄ–;Austrian Freedom Party;Male;0; 7;NA;NA; 50;1303;1040420;EP Candidate;MENF					
aximilian;FPÄ–;Austrian Freedom Party;Male;0; 8;maximiliankurz;84803032; 50;1303;1040420;EP Candidate;MENF					
drea;FPÄ–;Austrian Freedom Party;Female;0; 9;NA;NA; 50;1303;1040420;EP Candidate;MENF					
rin;FPÄ–;Austrian Freedom Party;Female;0; 10;NA;NA; 50;1303;1040420;EP Candidate;MENF					

If data have disclosure risks – access controls may be one solution

- ◊ “Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness”
- ◊ No tweet content, only IDs - to comply with Twitter Terms of Service
- ◊ Data accessible (by request) but not public because of no consent and reidentification risk
- ◊ Archived in SowiDataNet-*datorium*
 - ◊ Findable – Pfeffer, J. and Morstatter, F. (2016)
 - ◊ Preserved – DOI - (<http://dx.doi.org/10.7802/1166>)
 - ◊ Reproducible - Python scripts, tools, documentation
- ◊ [As open as possible, closed when necessary](#)

Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness

Cite as

URI	https://doi.org/10.7802/1166
Primary Researcher:	Pfeffer, Jürgen; Carnegie Mellon University Morstatter, Fred; Arizona State University
Publication Year:	2016
Availability:	Restricted Access
Other Contributors:	Zenk-Möltgen, Wolfgang ;GESIS - Leibniz Institute for the Social Sciences;Contact Person
Content	
Subject Area:	Information Science Mass Communication
Abstract:	This dataset consists of IDs of geotagged Twitter posts from within the United States. They are provided as files per day and state as well as per day and county. In addition, files containing the aggregated number of hashtags from these tweets are provided per day and state and per day and

Why archive data at all?

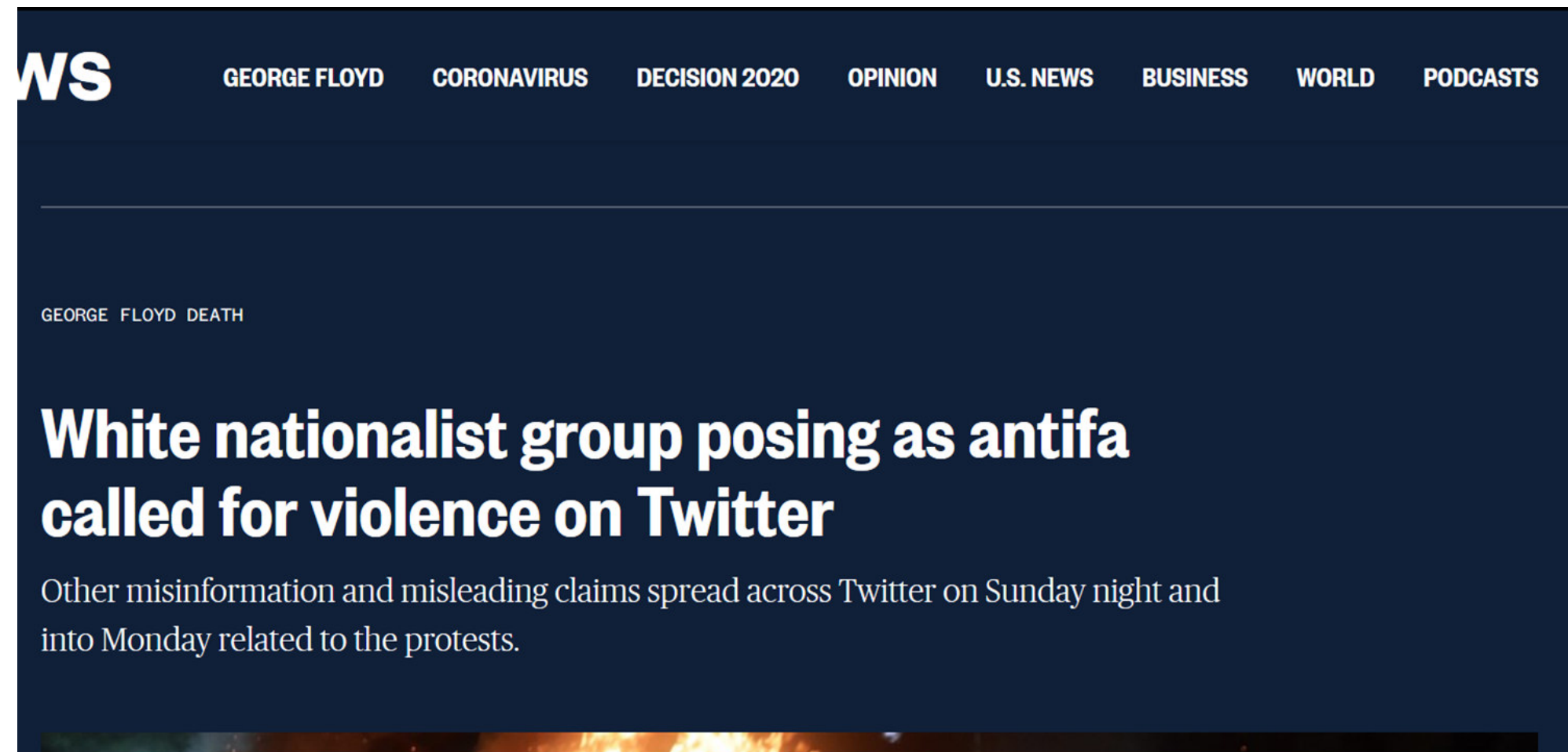
- ◊ Archives claim broad social responsibilities
 - ◊ Preservation – historical value
 - ◊ Reproducibility – integrity of data and methods
 - ◊ Open and FAIR data
- ◊ What (new?) responsibilities arise when archiving social media data?

Do archives have broader ethical responsibilities?

- ◊ A researcher wants to deposit data containing tweets from an account that was deleted by Twitter. She justifies the violation of Terms and Conditions on grounds of historic significance and public interest. Should you accept the data?
- ◊ A world famous researcher offers you gold star data that underpins a widely read policy article. Most of the data cannot be archived because of restrictions placed by the platform owner. Should you accept the data?

Which data have historical significance?

- ◇ Twitter account was a front, created impression that antifa was inciting violence at protests
- ◇ Clearly a violation of T&C to archive
- ◇ Is there a competing duty that overrides complying with T&C?

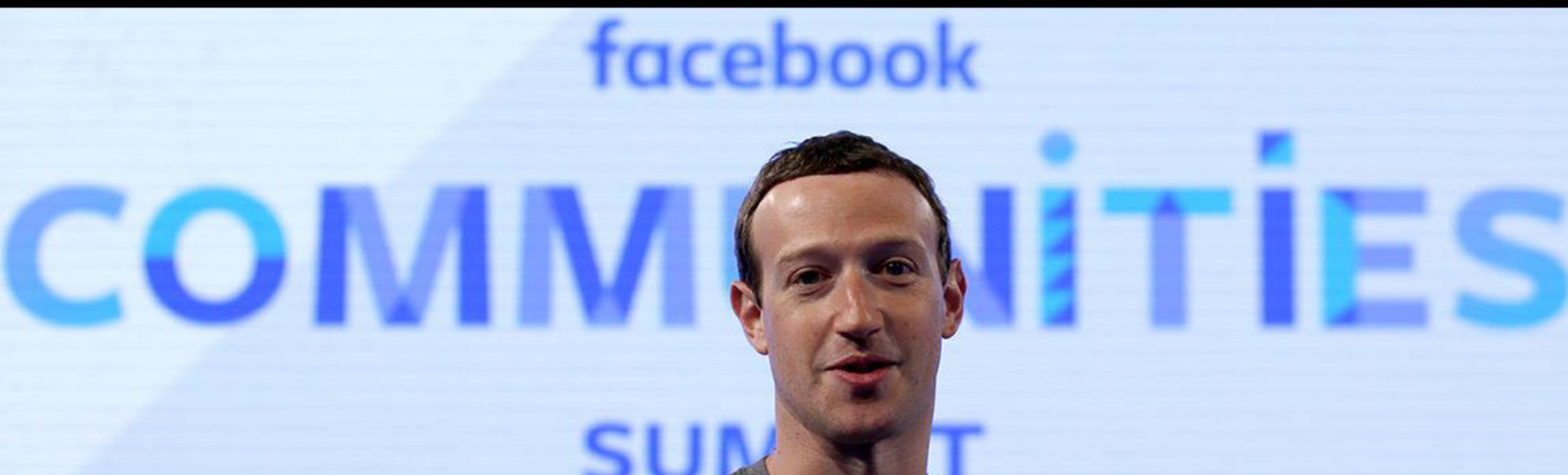


How far will we go to make data FAIR?

Facebook's next project: American inequality

A Stanford economist is using the company's vast store of personal data to study why so many in the U.S. are stuck in place economically.

By NANCY SCOLA | 02/19/2018 07:13 AM EST



- ◇ Raj Chetty (Prof, Harvard U) is doing unbelievably good work," said Harvard political scientist Robert Putnam "Mostly, it's because he's been able to get access to data that nobody else was able to get access to".

- ◇ <https://www.politico.com/story/2018/02/19/facebook-inequality-stanford-417093>

“Rules” have multiple interpretations...

- ◊ “At George Washington (GW) University Libraries, we (unofficially) interpreted this [3rd party sharing] to allow sharing Twitter datasets that we collected with anyone affiliated with GW (including students, faculty, and other researchers) and their collaborators. (What constitutes a “collaborator” is, of course, ambiguous.) If someone from outside GW contacts the library about a dataset, only the tweet ids are shared.”

- ◊ (Justin Littman, “Twitter’s Developer Policies for Researchers, Archivists, and Librarians” <https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2>)

Where do we stand?

Thank you

ElizabethLea.Bishop@geis.org

 cessda.eu

 @CESSDA_Data



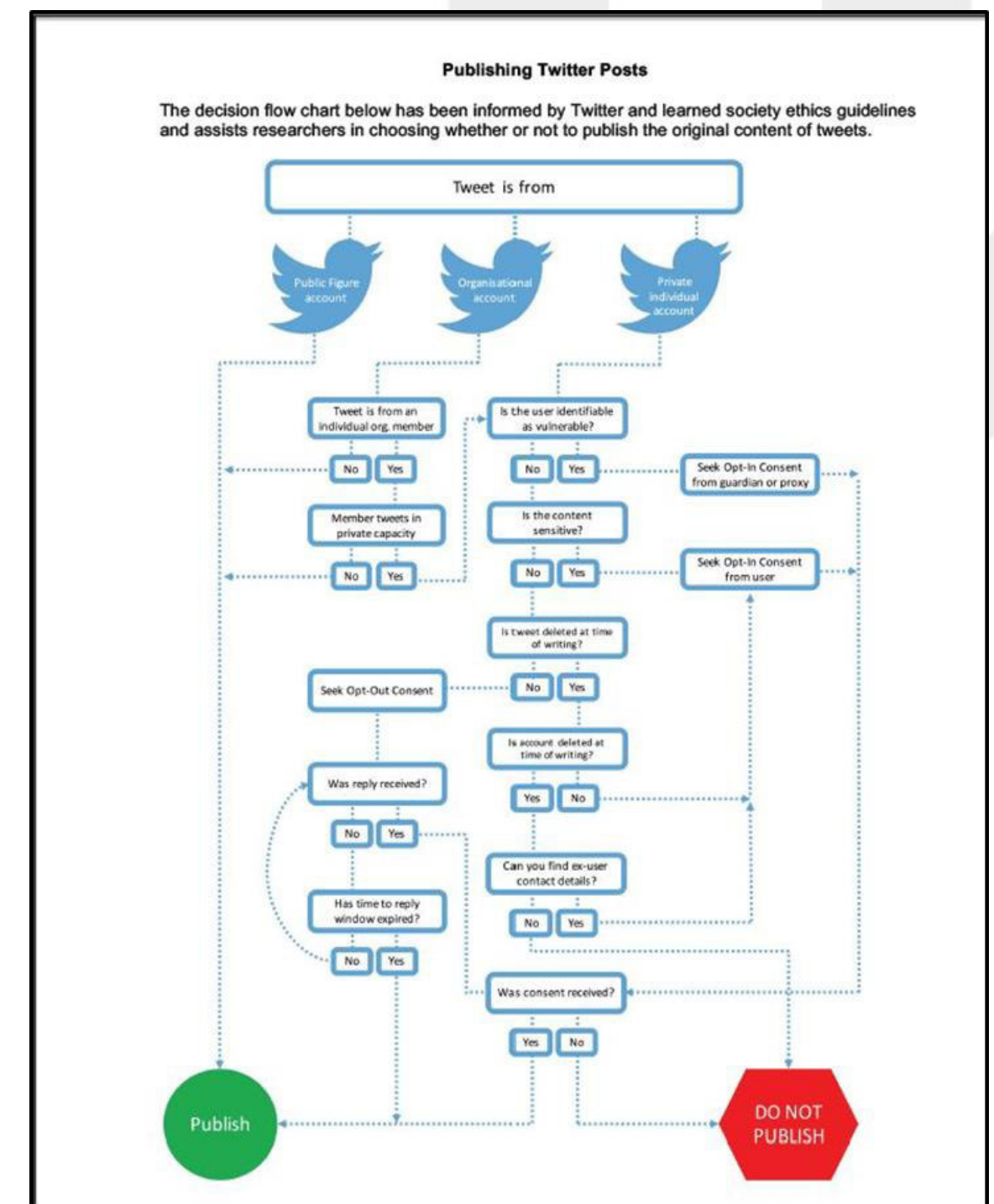
Licence: CC-BY 4.0

But I need an answer...

Informed consent	If using consent, do I need obtained consent for publishing (as well as for research)?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing.
	If consent is not possible, is publication ever permitted?	Yes. Key factors to assess are privacy, topic sensitivity, and subject vulnerability. Higher values increase the probability that consent is needed.
Private/public	My data are public; can I now publish without restrictions?	Possibly, yes. But ethical considerations remain, even for "public data".
	How public or private is the source (forum, etc.) from which I collected the data?	It is necessary to assess the "publicness" of the setting. Requirements to register, presence of a moderator, password protection, etc. all suggest some intent toward private communication. Open access, institutional accounts, and broadcast messages all suggest more public intentions.
Vulnerability	Can I publish content from vulnerable participants (e.g. children, elderly)?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Sensitive data or topics	Are the data about sensitive topics (e.g. health, religion, political views, sexuality etc.)? If yes, have I considered if the potential benefits of this research offset the additional risks?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Access restrictions	My publisher insists that I deposit my data, but my data have disclosure risks. What are my options?	Most publishers permit an alternative, a "data access statement" explaining restrictions. Some repositories provide gradations of control that may enable regulated access to data.
If publication/sharing not possible	What should I do if the platform [Twitter, Facebook, etc.] does not allow me to publish any data?	Check any Terms and Conditions.
	My data simply cannot be published.	Consider how you will handle replication requests.
D. Sharing		
	Questions	Guidance
Research Ethics Approval	Are my data sharing plans clearly described in my Ethical Review application?	Data sharing intentions should be made clear in the ethics application. It may be possible to provide the name of a social media researcher to your REC if the committee does not have such expertise.
Informed consent	If using consent, do I need consent for data sharing?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing.
	Is broad consent an accepted ethical practice when data may be reused for purposes other than the primary one, or even for unknown purposes?	Broad consent is still used in domains such as biobanking. GDPR (Recital 33) allows for consent even when detailed purposes are not known, subject to "recognized ethical standards".
	If consent is not possible, is data sharing ever permitted?	Yes. In addition to topic sensitivity, subject vulnerability and privacy, other factors to consider are: is the research impossible with consent, could the research be done with another method, and are there compensating benefits from doing the research?
Private/public	My data are public; can I now archive and share what I want?	Possibly, but ethical duties may prevent sharing, e.g., if a participant has consented, but the researcher believes she does not fully understand the risks of data sharing. (See legal section for other restrictions)

Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation
 Matthew L Williams, Pete Burnap, Luke Sloan
Sociology, First Published May 26, 2017
<https://doi.org/10.1177/0038038517708140>

Appendix A: Short guide on legal and ethical issues for the researcher to consider when using social media for research
 WP6-D3 Report (not guidelines)
https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies..._FINAL.pdf



Archiving Qualitative Social Media Data

Sebastian Karcher (Qualitative Data Repository)

CESSDA Webinar

“Archiving Social Media Data: Challenges and Proposed Solutions”

June 4, 2020



This work is licensed under a
[Creative Commons Attribution
4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Distinctive Characteristics of *Qualitative* Social Media Data

- Manually collected:
 - Wide range of sources within single project
 - Typically no use of API
- Small(-ish) n: Individual assessment possible, e.g.
 - Creators' intent
 - Originality
 - Copyright / Fair use
 - Consent

Case Study: Clarke (2018)

Volume 16, Issue 3 September 2018 , pp. 617-633

When Do the Dispossessed Protest? Informal Leadership and Mobilization in Syrian Refugee Camps

Killian Clarke

<https://doi.org/10.1017/S1537592718001020> Published online: 21 August 2018

- Supplementary Data to article in *Perspectives on Politics*,
- Why and when do refugees in camps/settlements protest? Common in Jordan, but not in Turkey and Lebanon
- UNHCR Reports, Social Media “Event Data”, Interviews (not shared)

Clarke, Killian B. 2018. "Data for: When do the dispossessed protest? Informal leadership and mobilization in Syrian refugee camps". Qualitative Data Repository. <https://doi.org/10.5064/F6CN723S>

Clarke: Event Data Challenges (Ethics, Copyright, Preservation)



LBCI Lebanon News EN

@LBCI_News_EN

Follow

Syrian nationals hold demonstration in support of #Assad in #Zahle | lbcgroup.tv/news/155969/LB... | #Syria #Lebanon

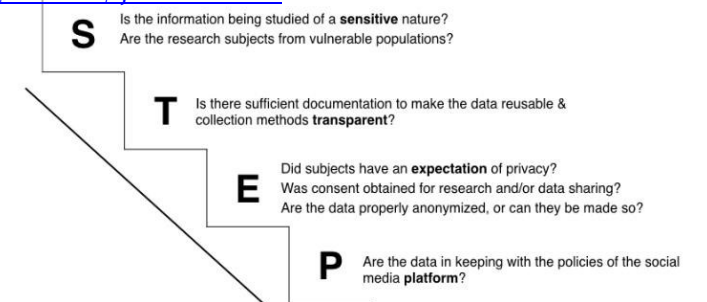
3:49 AM - 18 May 2014

Sharing Selves: Developing an Ethical Framework for Curating Social Media Data

Sara Mannheimer
Montana State University

Elizabeth A. Hull
Dryad Digital Repository

<https://doi.org/10.2218/ijdc.v12i2.518>



May 18, 2014	Zahle	Protest	http://www.twitter.com/LBCI_News_EN/status/467950093274865664	Good	https://www.lbcgroup.tv/news/155969/LBCI-News ... #Syria #Lebanon
--------------	-------	---------	---	------	--

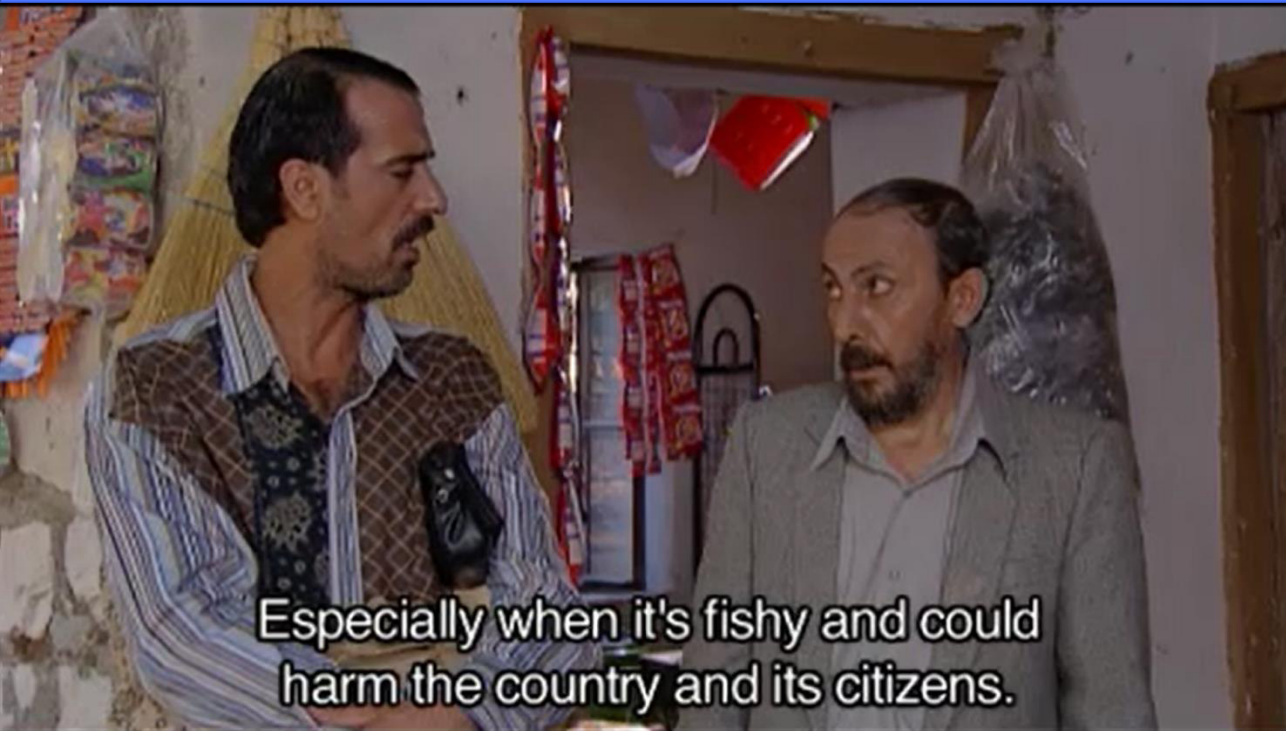
<https://www.lbcgroup.tv/news/155969/LBCI-News>

<https://perma.cc/HH6H-3MAY>

Tools: perma.cc & Internet Archive

- Both provide archiving capabilities for web content
- IA
 - Free to use
 - Large existing archive (WayBackMachine)
 - Awkward API for archiving
- Perma.cc
 - Archiving requires payment of membership through library
 - Archiving via API
 - Screenshots and regular WARC files

Case: Syrian Civil War



Wedeen, Lisa. 2019. "Data for: Authoritarian apprehensions: Ideology, judgment, and mourning in Syria." Qualitative Data Repository.

"What led a sizable part of the citizenry to stick by the regime through one atrocity after another? What happens to political judgment in a context of pervasive misinformation? And what might the Syrian example suggest about how authoritarian leaders exploit digital media to create uncertainty, political impasses, and fractures among their citizenries?"

- Hundreds of online sources used throughout text
- Many of them video (Youtube, Vimeo, etc.)
- Additionally, videos from Syrian TV licensed for copyright

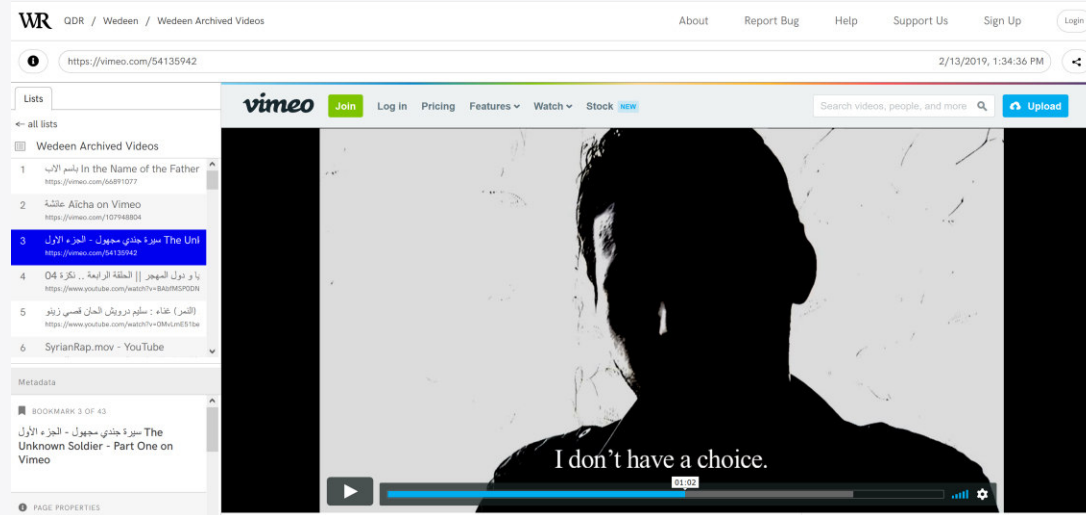
Tool: archivr

- R-tool developed by QDR
- Scans text (or spreadsheet) extracts URLs
- Option to archive using IA or perma.cc
- Returns list with original URLs and archived URLs (or error)
- <https://github.com/QualitativeDataRepository/archivr/>

```
> df <- view_archiv.fromUrl("https://www.cessda.eu/News-Events/News/CESSDA/Tour-of-CESSDA-The-Portuguese-Archive-of-Social-Information")
```

	url	available	wayback_url	timestamp
1	http://datacatalogue.cessda.eu	TRUE	http://web.archive.org/web/20191101135155/https://dataca...	20191101135155
2	https://www.europeansocialsurvey.org/	TRUE	http://web.archive.org/web/20200602030513/http://www.e...	20200602030513
3	http://capp.iscsp.ulisboa.pt/en/projects/ongoing-projects/2...	TRUE	http://web.archive.org/web/20191125090104/http://capp.is...	20191125090104
4	https://www.fct.pt/	TRUE	http://web.archive.org/web/20200513133038/https://www.f...	20200513133038
5	https://www.cessda.eu/About/Projects/Past-projects/CESSD...	TRUE	http://web.archive.org/web/20190605124231/https://www.c...	20190605124231
6	https://datacatalogue.cessda.eu/	TRUE	http://web.archive.org/web/20191101135155/https://dataca...	20191101135155
7	https://datacatalogue.cessda.eu/?publisher.publisher[0]=Ar...	FALSE	url not found	unknown
8	https://www.coretrustseal.org/	TRUE	http://web.archive.org/web/20200526174744/https://www.c...	20200526174744
9	https://www.sshopencloud.eu/news/developing-sshoc-data...	FALSE	url not found	unknown
10	https://www.cessda.eu/Tools-Services	TRUE	http://web.archive.org/web/20200403164118/https://www.c...	20200403164118
11	https://vocabularies.cessda.eu/	TRUE	http://web.archive.org/web/20200602120005/https://vocab...	20200602120005

Tool: Webrecorder.io



“Webrecorder is both a tool to create high-fidelity, interactive recordings of any web site you browse and a platform to make those recordings accessible.”

- Used by QDR to make videos and other non-static formats available
- Allows for publish sharing of collections (similar to IA/perma.cc)

A Very Recent Example: Individual Consent

From IASSIST List:

“A post-doc here wants to download posts and comments from a private Facebook group (...) The group was created by and is managed by some faculty here, so [the administrators and members of the group are on-board with the project and with using the posts/comments for research purposes.](#)”

General Lessons

Sharing Qualitative Social Media Data

- Similar challenges to sharing larger-scale social media data
- Given smaller number of items, individual curation & checks are feasible, allowing for more sharing
- Trade-off to “outsourcing” archiving to services such as perma.cc and webrecorder.io, but both the technical and legal benefits outweigh the risks
- Especially for larger scale efforts, further tools are available from Documenting the Now: <https://www.docnow.io/>

Questions? Comments?

Please stay in touch:

<https://qdr.syr.edu>



@adam42smith (Sebastian)

@qdrepository (QDR)

Email:

skarcher@syr.edu

qdr@syr.edu

QDR
**The Qualitative
Data Repository**

Archiving Social Media Data – Challenges and Proposed Solutions

Legal Issues

Oliver Watteler (GESIS)

*CESSDA Webinar Archiving Social Media Data –
Challenges and Proposed Solutions (4th of June 2020)*

 cessda.eu

 @CESSDA_Data



Licence: CC-BY 4.0

Archiving social science research data

- ◊ View of a repository / research data center / data archive
- ◊ Archiving = preservation + documentation + publication (e.g. CoreTrustSeal)
- ◊ Pre-Ingest (> archive agreement):
 - ◊ Clarify legal basis of data collection and thus rights to the data
 - ◊ Clarify conditions for re-use
- ◊ Ingest - ingest check
 - ◊ Check data quality
 - ◊ Check content of data and documentation for legal issues
 - ◊ Information reduction if necessary
 - ◊ Restrict access if necessary

What are Social Media data?

- ◊ Social media data - attributes (Obar, Wildman 2015):
 - ◊ Internet-based applications
 - ◊ User-generated content
 - ◊ User-specific profiles for individuals and groups
 - ◊ Social networking by connecting profiles of different users
- ◊ Which (most prominent) rights are involved in the use of Social Media platforms?
 - ◊ Contractual agreement between user and platform provider ('Terms of Service')
 - ◊ Data protection for personal information
 - ◊ Intellectual property rights (IPR) for content like photos, videos, audios, (creative) texts
 - ◊ Database rights

How are Social Media data collected?

- ◊ Research interests?
 - ◊ As diverse as social research (Van Osch, Coursaris 2015)
- ◊ Which ways of collection are there (Breuer et al. 2020)?
 - ◊ Web scraping – problem: restricted access to some information behind login
 - ◊ Application Programming Interfaces (API) – bound by technical limitations
 - ◊ Cooperating with platform provider (“privileged access”)
 - ◊ Purchase the data
- ◊ What are the legal bases for data collection?
 - ◊ Application of a law (rather not in our case)
 - ◊ Informed consent (scraping with consent, “data donation”, tracing with sensors ...)
 - ◊ Freedom of research (not equally established internationally; Santosuosso 2012)?
 - ◊ **Agreement to ‘Terms of Service’** (usage agreement, purchasing contract)
 - ◊ Circumvention a „gray area“ (Halavais 2019, Franzke et al. 2020)
 - ◊ Illegal techniques applied? (“hacking”, deceit, fraud, ...)

What is supposed to be archived?

What data are we looking at?	What are the legal impediments?	Possible challenges
Personal data (e.g. names if no alias or nickname, personal information) Also from other individuals!	Data Protection Legislation	Breach of confidentiality, breach of privacy, re-identification (also others)
Images / videos / sound recordings	IPR, Data Protection Legislation	Copyright violation, breach of individual personal rights
Texts (depending on the level of creativity)	IPR	Copyright violation
Parts of a database (platform providers)	Database Protection	Database creator's rights violation
Technical metadata (e.g. time stamps, IP ranges, ...)	?	Some technical metadata might enable disclosure of information (e.g. preferences), or lead to 'profiling' (linking various social media profiles)

- ◊ **Main problem:** data might have been collected without neither consent by the 'data subjects' (GDPR term) nor consent by the platform providers (violation of 'Terms of Service')

Means of archives to meet the challenges?

- ◊ In the ideal case: **all rights are clarified** - in the real world: **often rights are not clarified**
- ◊ “Classical instruments” of data archiving on data level (pro’s and con’s):

	Activity	Pro’s	Con’s
A	Data Protection <ul style="list-style-type: none"> • Information reduction (pseudonymization / anonymization) • Aggregation 	<ul style="list-style-type: none"> • No need to protect anonymized data 	<ul style="list-style-type: none"> • Anonymization hardly possible (Narayanan, Felten 2014) • Information reduction reduces value of data
B	Protection of copyright <ul style="list-style-type: none"> • Delete copyright protected material • Keep content if possible (e.g. text korpus) 	<ul style="list-style-type: none"> • Non copyright-protected information can be published (without personal information under A) 	<ul style="list-style-type: none"> • Information reduction reduces value of data
C	Protection of databases <ul style="list-style-type: none"> • Only archive (part of) content of database 	<ul style="list-style-type: none"> • Data („facts“) per se (if not under A or B) often not protectable 	<ul style="list-style-type: none"> • Access to data more difficult

- ◊ If data was collected legally and purpose of use remains public interest (esp. UK), science or statistics: restrict access (for ‘person-related’ data, see GDPR)
- ◊ Other tools might be transferrable - e.g. ‘fabrication’ (≈ ‘synthetic data’ in Official Statistics)
- ◊ Repositories’ answer: need for new tools to match Big Data’s 3 v’s (volume, velocity, variety – sometimes also veracity); solutions **in the making** (Kinder-Kurlanda et al. 2017; Breuer et al. 2020)

Thank you for your attention

oliver.watteler@geis.org

 cessda.eu

 @CESSDA_Data



Licence: CC-BY 4.0

References

- ◊ AAPOR (2015): AAPOR Big Data Task Force, AAPOR Report on Big Data, <https://www.aapor.org/Education-Resources/Reports/Big-Data.aspx>; last access: 01.06.2020
- ◊ Breuer et al. (2020): Johannes Breuer, Libby Bishop, Katharina E. Kinder-Kurlanda (Forthcoming), The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships, in: *New Media & Society*
- ◊ CoreTrustSeal (2018): CoreTrustSeal, Core Trustworthy Data. Repositories Extended Guidance, 2018, <https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf>; last access: 29.05.2020
- ◊ Corti et al. (2020): Louise Corti, Libby Bishop, Veerle Van den Eynden and Scott Summers, Working with Big and Novel Data, in: Louise Corti et al., *Managing and Sharing Research Data. A Guide to Good Practise*, SAGE, 308-332
- ◊ franzke et al. (2020): aline shakti franzke, et al., Internet Research: Ethical Guidelines 3.0, <https://aoir.org/reports/ethics3.pdf>; last access: 29.05.2020
- ◊ Halavais (2019): Alexander Halavais, Overcoming terms of service: a proposal for ethical distributed research, in: *Information, Communication & Society*, 22(11), 1567-1581, <https://doi.org/10.1080/1369118X.2019.1627386>; last access: 29.05.2020
- ◊ Kinder-Kurlanda et al. (2017): Katharina Kinder-Kurlanda et al., Archiving Information from Geotagged Tweets to Promote Reproducibility and Comparability in Social Media Research, in: *Big Data & Society*, 4 (2), 1-14, <http://dx.doi.org/10.1177/2053951717736336>; last access: 02.06.2020
- ◊ Lynch (2017): Clifford Lynch, Stewardship in the "Age of Algorithms", in: *First Monday*, 22(12), <https://doi.org/10.5210/fm.v22i12.8097>; last access: 29/05/2020
- ◊ Narayanan, Felten (2014): Arvind Narayanan, Edward W. Felten, No silver bullet: De-identification still doesn't work, Princeton, <https://www.cs.princeton.edu/~arvindn/publications/no-silver-bullet-de-identification.pdf>; last access: 29.05.2020
- ◊ Obar, Wildman (2015): Social media definition and the governance challenge: An introduction to the special issue, in: *Telecommunications Policy*, 39(9), 745-750, <https://doi.org/10.1016/j.telpol.2015.07.014>; last access: 29/05/2020
- ◊ Santosuosso (2012): Amedeo Santosuosso, Freedom of Research and Constitutional Law. Some Critical Points, in: Simona Giordano et al. (Eds.), *Scientific Freedom: An Anthology on Freedom of Scientific Research*, London, 73-82, <https://www.bloomsburycollections.com/book/scientific-freedom-an-anthology-on-freedom-of-scientific-research/ch6-freedom-of-research-and-constitutional-law>; last access: 29/05/2020
- ◊ Van Osch, Coursaris (2015): Wietske Van Osch, Constantinos K. Coursaris, A Meta-Analysis of Theories and Topics in Social Media Research, 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, 2015, pp. 1668-1675, <https://ieeexplore.ieee.org/document/7070011>; last access: 29/05/2020

Archiving Social Media Data

Challenges and Proposed Solutions

Roundtable Discussion



Questions

@ 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?

@ 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)

@ Janez Štebe:

Is there minimum requirement for data to meet each of the FAIR principles? For example 80% or higher?

Is it legal to use images/videos shared on the social media platform? For example, profile pictures, screenshots showing tweet or Facebook posts

What if researchers do not know which
Terms of Service they agreed to?
What if they scraped the data?

Where/how do copyright and 'legal restrictions' impact archiving social media data?

Who should archive social media data?

For how long is the commitment?

Where is the funding to curate the
collections coming from?

follow up question to my previous question, for a research, if we are working with 1000 twitter profile, It is not possible to ask everyone and take permission to use their profile picture and name. In this case, can we stop bothering about about IPR and publish the results? Or what should we do?

This is a question for everyone not from GESIS :) 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 just 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 mentioned needed to be very actively recruited (eventually by making it a paper project:

<<https://journals.sagepub.com/doi/pdf/10.1177/2053951717736336>>) - so considerable effort from various GESIS archivists was required.

« everybody » , what type of standards
do you use to implement metadata ?
DDI, RDF , (Disco) ?

Thank You!

 cessda.eu

 @CESSDA_Data



Licence: CC-BY 4.0