ADP - SOCIAL SCIENCE DATA ARCHIVES

Analyze data! Deposit study! Promote science!

seriss SYNERGIES FOR EUROPE'S RESEARCH INFRASTRUCTURES IN THE SOCIAL SCIENCES

FAIRness of (Linked) Social Media Data

Janez Štebe, ADP [Arhiv družboslovnih podatkov], University of Ljubljana

ESRA 8th Conference Zagreb, Croatia, from 15th to 19th July

Slovenian Social Science Data Archives (*ADP-Arhiv Družboslovnih Podatkov*)

- Founded in 1997
- Slovenian national research data centre for social sciences
  - 600 social science studies data accessible in a data catalogue + 150 metadata only
  - Mainly survey data (from 1960's on), few qualitative, social networks and social media
- member of CESSDA ERIC
- obtained CoreTrustSeal in beginning of 2018
- involved in EU, CESSDA and national projects

CORE TRUST SEAL

ADP Social Science Data Archives
http://www.adp.fdv.uni-lj.si/eng/
CTS Certification 2017-2019

ADP involvement in SERISS project

- Linked third party CESSDA partner in SERISS

- Involved in Task 1 and 3, WP6: New forms of data – legal, ethical and quality issues
- Task 1 address: Legal and ethical etc. challenges of Social Media (SM) Data
- Results: Workshops, reports, guides
- Tasks partners: NSD (WP and T1 Lead), UKDA (T3 Lead), ČSDA, ADP, (GESIS)

- In this presentation : Special attention about the data service aspects of dealing with SM data:
    - Appraisal and selection, Curation, Archiving, Sharing, Disseminating and Re-using SM data

Why is discussing about sharing SM data important?

- SM research: becomes established source of data for research (complementing, substituting other sources/methods, including surveys- see Kleiner 2015, Callegaro 2018)
- SM data can be reused for the benefits of new research
- Journals, funders request data sharing for transparency and reproducibility of research (under FAIR principles)
- Data sharing principles are (still) in flux (we need to discuss and re-evaluate, establish standards) (Mannheimer and Hull 2017)
- CESSDA: only a handful of SP is accepting/ curating SM data (see appendix of SERISS 6.3 report)

Issues about (sharing) SM data

- Compared to survey data: this are ‚organic' data, found and not generated with research purpose in mind
- Issues about data quality (Japec et al. 2015; Callegaro 2018)
- Complexity of Big Data (size, structure,…)
- 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)

Paradigmatic cases of (non)sharing SM data

The following categories of cases:

- **Establishing best practice**: Transparent arguments about principles and decisions taken that lead to archived and accessible SM data (Kinder-Kurlanda et al. 2017)
- **Typical** for certain group of cases: e.g. sharing for replication purpose in a self-archiving repository by disclaimer (Chukwuemeka and Abdul 2017)
- **Exceptional**: resolving conflicts of principles showing an exception from the typical rules (Mannheimer, Hull 2017)
- **Problematic**: reach attention as cases that cross the boundary of acceptable (e.g. public vs. private consideration in **Tastes, Ties, and Time Dataverse**)

Range of solutions we can observe:

- Users / depositors responsibility vs. curator responsibility
- Comparison of general repositories vs. disciplinary data centres

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*.‘
- Attributes of metadata and data to qualify as Findable, Accessible, Interoperable and Reusable

Discussions and misconceptions about FAIR :

- Not new, just make it more visible some of the long standing principles around open research data (OECD, Royal society, …)
- Oriented towards Machine- computer automatic assessment
- Overlap with some other principles: regarding data citation (Force11), Core Trust Seal (more focus on repositories, while FAIR is data centric)
- FAIR is not equal Open (there are various degrees of openness that are legitimate reasons for not completely open depending on data characteristics; there are degrees of FAIRness that are appropriate for the type of data in specific case)

Selection from existing FAIR maturity metrics for the purpose of evaluating SM data sharing

- Different emphasis on some of the aspects depending on discipline established requirements, type of data and purpose of evaluation…

- including a discussion about what is perhaps missing from usual FAIR criteria:
    - data quality assessment,
    - and costs of keeping and curating data

What we will do:

- elaborate FAIR criteria and Map to Attributes of cases of SM data shared (and repositories)

Sources of existing metrics for FAIR

ANDS-NECTAR-RDS-FAIR data assessment tool	ARDC
DANS-Fairdata	DANS
DANS-Fair enough?	DANS
The CSIRO 5-star Data Rating tool	CSIRO
FAIR Metrics Questionnaire	The FAIR Metrics Group
Stewardship Maturity Mix	NOAA's CICS-NC, NOAA's NCDC
FAIR Evaluator	GO FAIR, LUMC CBGP, IDS, RDA FAIRsharing, IQSS
Data Stewardship Wizard	ELIXIR NL/CZ
Checklist for Evaluation of Dataset Fitness for Use	Assessment of Data Fitness for Use WG (WDS/RDA)
RDA-SHARC Evaluation	SHARC IG (RDA)
WMO-Wide Stewardship Maturity Matrix for Climate Data	The SMM-CD WG
Data Use and Services Maturity Matrix	The MM-Serv WG

Taken from: RDA Fair Maturity Assessment

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier

Possible criteria (relevant for SM data, but could be quite general) selected from the list of existing proposals (see RDA Fair Maturity Assessment):

- *Citation exists, including authorship, year, comprehensive title, persistent identifier (e.g. DOI)*
- *Persistent identification of the dataset and related work (related literature and data, authors, projects, terms)*

***Related SM Data** issues:*

Are the data, literature and code and project etc. PID-s present? Or linked in a persistent way.

Is it possible to cite data using required elements? Is the data cited by the related article?

Example cases

Data deposited ✓ in the disciplinary data archives:

- Kaczmirek, Lars; Mayr, Philipp (2015): German Bundestag Elections 2013: Twitter usage by electoral candidates. GESIS Data Archive, Cologne. ZA5973 Data file Version 1.0.0, http://dx.doi.org/10.4232/1.12319 ✓

SM data linked to survey data not ✗(yet) archived, related to:

- Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2019). Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving. Journal of Empirical Research on Human Research Ethics. https://doi.org/10.1177/1556264619853447 ✗

„(…)none have gone through a formal archiving process and been accessed by researchers working independently of the original research team. Although we have identified what we think may be key issues and how they may be overcome, it will only be through actually archiving and providing access to these data that we might fully understand the challenges and whether or not the measures we have outlined will address them."(Sloan et al. 2019: 10)

Remaining F… aspects

F2. Is data described with rich metadata

F3. Metadata clearly and explicitly include the identifier of the data it describes

SM (meta)data criteria:

- Extent of descriptive information in metadata
- Standard disciplinary metadata used, cross walked to general repository metadata systems
- Searchable on the institutional, disciplinary, and/or general catalogue…

SM Cases:

- Informal sharing, sharing 'upon request', sharing on website, GitHub: ✗
    - that lack the availability of search, insufficient documentation, and poor response to request access (mentioned in Weller and Kinder-Kurlanda 2016)

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol

A1.1 the protocol is open, free, and universally implementable

Example criteria:
Data not available publicly; Person-to-person contact needed.
Basic online services available for data access (e.g. FTP/HTTP direct download).
Non-standard data services.
Standard-based interoperability data services.
Previous + Full capability of sub-setting, aggregation and visualization.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

Example criteria:
Please provide a IRI that resolves to a description of the process to obtain access to restricted content In case of a non legal restricted access, is the restriction properly justified by the researcher ?

Example cases

SM (meta)data criteria:

- Establishing existence of metadata describing access conditions…and reasons for not completely open access…

SM Cases:

GESIS example:

- Kinder-Kurlanda et al. (2017) arguing that it‘s possible to re-identify individuals based on rehydrated twits.
- No consent was obtained.
- Location can be reconstructed

As a consequence:

‚Each access request is decided on individually based on the information provided in the application (e.g., research topic, methods, etc.)'. ✓ (Kinder-Kurlanda et al. 2017)

A2. metadata are accessible, even when the data are no longer available

Example criteria:
Metadata persistence policy / guarantee
Comment: based on repository having a plan (CTS)…

SM (meta)data *test:*

Access to metadata and code (even if data is not publicly accessible)?

E.g. ID of Twits: full metadata is searchable and accessible… even if 'deidentified' /modified data only is accessible.

Problematic:

- *Persistence of content (if post behind the ID – lost),*

Example cases

NSD full metadata without data:

Social Media During and After the Terrorist Attacks (Hornmoen, H., 2017). https://doi.org/10.18712/nsd-nsd2434-v1

Availability Status / Restrictions

– *For further information please contact the principal investigator.*

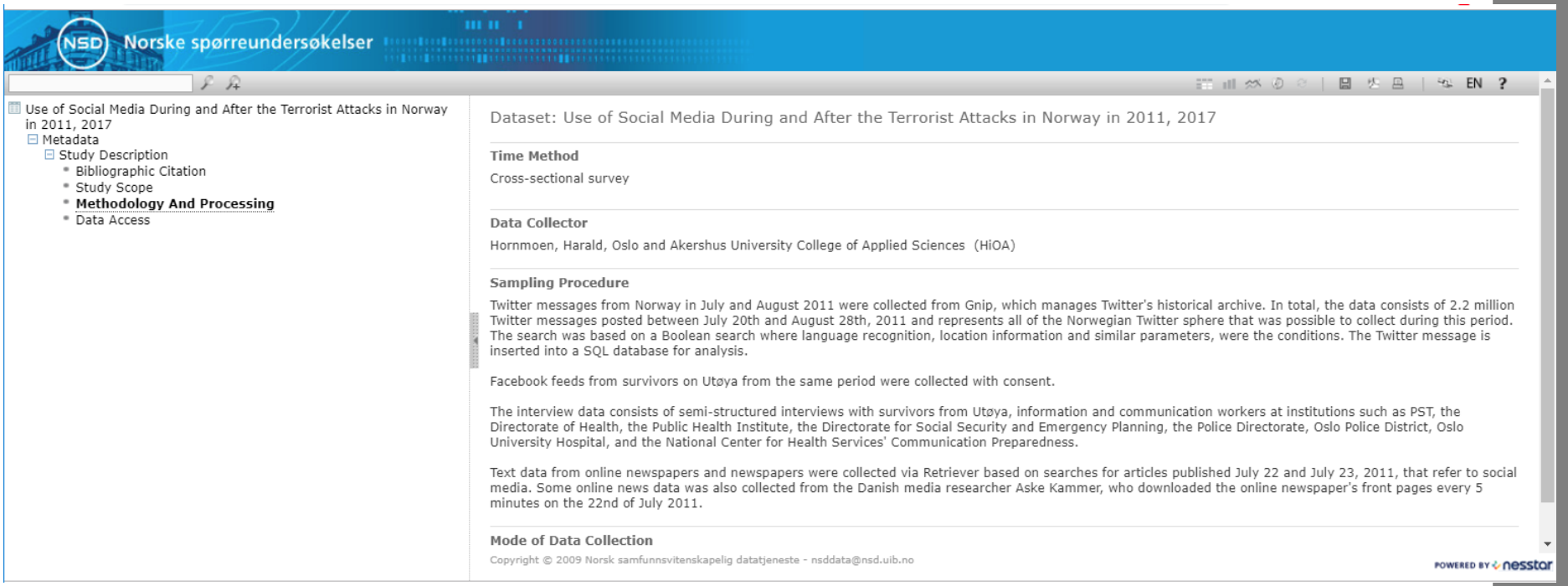
Tastes, Ties, and Time Dataverse:
https://dataverse.harvard.edu/dataverse/t3)

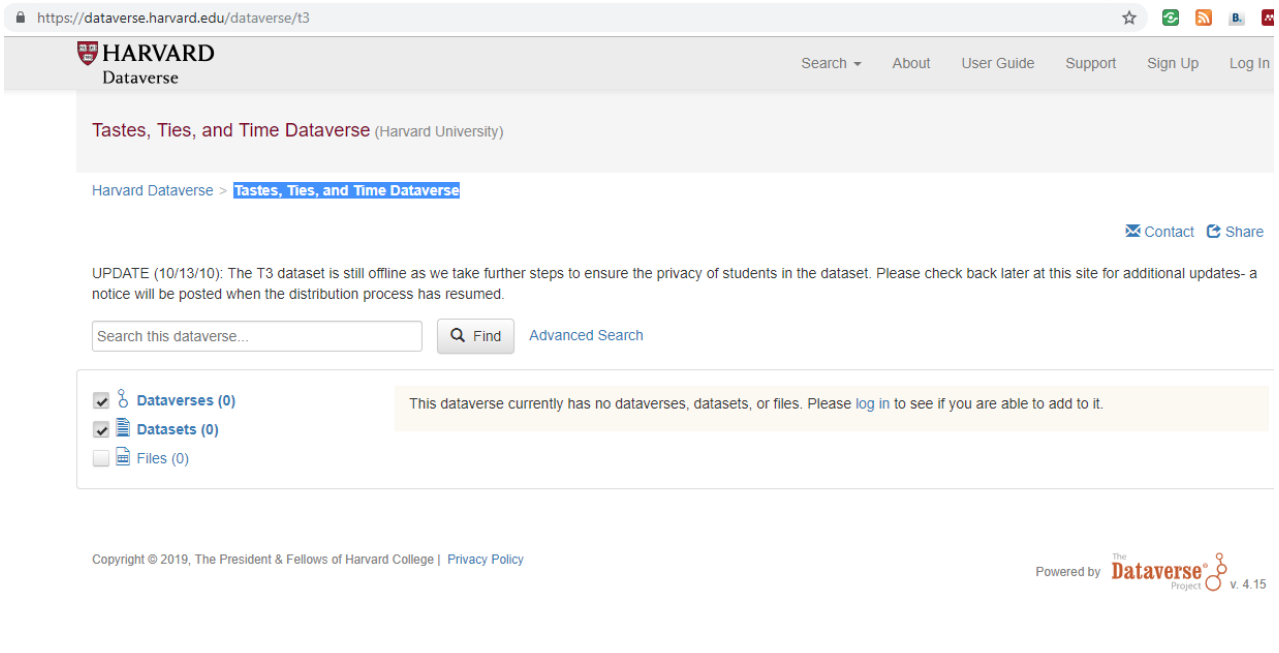
DataVerse landing page: denoting non-existence…

– Fulfilment of only one of requirements: Persistent DataVerse identifier (internet location) that lead to a note:

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.

- No metadata about past existing data ✗
- If the comprehensive metadata would exist, this could serve as an intermediate substitute for data


Norske spørreundersøkelser
EN
Use of Social Media During and After the Terrorist Attacks in Norway in 2011, 2017
Metadata
Study Description
Bibliographic Citation
Study Scope
Methodology And Processing
Data Access
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
Copyright © 2009 Norsk samfunnsvitenskapelig datatjeneste - nsddata@nsd.uib.no
POWERED BY nesstar


https://dataverse.harvard.edu/dataverse/t3
HARVARD Dataverse
Search
About
User Guide
Support
Sign Up
Log In
Tastes, Ties, and Time Dataverse (Harvard University)
Harvard Dataverse > Tastes, Ties, and Time Dataverse
Contact
Share
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 this dataverse...
Find
Advanced Search
Dataverses (0)
Datasets (0)
Files (0)
This dataverse currently has no dataverses, datasets, or files. Please log in to see if you are able to add to it.
Copyright © 2019, The President & Fellows of Harvard College | Privacy Policy
Powered by The Dataverse Project v. 4.15

To be Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles

SM (meta)data criteria:

- added value processing: standard variables/attributes annotation…

SM case:

Ljubešić, Nikola; Erjavec, Tomaž and Fišer, Darja, 2017, *Twitter corpus Janes-Tweet 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1142.

– linguistic annotation (normalised and lemmatised, TEI encoding), gender, author type, sentiment

I3. (meta)data include qualified references to other (meta)data

Not specific to SM data:

Reach context information for related information about project, publications etc...

global journal related repositories are favourites there…

But specific is:

links to code, perhaps anonymised versions of data, … in particular when access to data is limited.

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes

Example criteria:
Database has users guide
Is there extensive metadata and rich additional documentation available (for others to understand and reuse your data)?
Content of the dataset agrees with description of the dataset content
Which relevant actions have been undertaken by the researcher to enhance the data reuse potential
Does the researcher provide information on methods and tools that permit the understanding, integrity, value and readability of data intended to be kept on the long-term ? (e.g. versioning, archival and long term reuse issue for protocols, softwares, required methods and contexts to create, read and understand data)

Cases

SM Case of extended documentation and data use suggestions:

International Conference on Web and Social Media (ICWSM) dedicated **Data papers section**. E.g.:

- Brena, 2019a: News Sharing Users Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions

- The data referred to is available at one of the dedicated repositories, in this case it is at the Harvard Dataverse (Brena, 2019). The code is published on GitHub.

SM Case: Hemphill et al (2018): Developing a Social Media Archive at ICPSR:

ICPSR Jupiter notebook feature plan of SM archive to enhance the replicability…

R1.1. (meta)data are released with a clear and accessible data usage license

Example criteria:
Terms of usage (licenses, other conditions of reuse, data protection, ethical issues)
Do the data reuse control and data sharing arrangements meet the data protection and local/national ethics requirements?
Legal reuse restriction properly justified?

Example case

Chukwuemeka , David and **Abdul, Adeniyi** (2017) GE_Readme.docx:

It is noted that we only upload a subset of the collected Tweets (1499,999 Tweets) to comply with Clause F.2 of the Twitter Developer Policy (effective June 18 2017).

Disclaimer

The published Data are for non-commercial research purposes only, and the use of the data is subject to all of the other provisions and rules of the Twitter Developer Policy and Agreement.

Data contained within each tweet sample are copyright by the Tweet's authors and are subject to all copyright law. The creator of the dataset makes no assertion of rights related to the content of each tweet. The dataset is provided 'as is' without warranty or any representation of accuracy, timeliness or completeness.

Please follow Twitter Term of Service and related policies when accessing the above files and using the data contained in these files. We take no responsibility for any inappropriate/illegal use of the data by any third party.

R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Description of methods used to create this dataset are appropriate for the context and discipline

SM Case typical observation:

Hemphill et al (2019): How can we save social media data?:

- *‚Researchers then rarely describe the particulars of those collection methods or the transformations they perform on the data to prepare it for analysis. The inability to judge the quality or understand the provenance of a single research group's effort presents additional challenges for other research groups to reuse the data.'*

Cost considerations

Not all data need to be shared /

Not all data need to be open (only as open as it‘s reasonable and fair)

- Creator (author) or repository (data archive): the data archive intensity of curation follows the appraisal and selection criteria of collection: only reference data sets deserve highest intensive level of curation
- Trade of between usability and access: anonymization can be time consuming and costly; or even not possible for SM content (and related survey data);
- Controlled access perhaps better solution? (but views differ)

Appraisal and Selection for new type of data

- Repositories poses general-purpose disciplinary knowledge of the data and curation best practice:
- Compared to researchers specific knowledge of study design, data collection and analysis.

Conclusion

FAIR is less a matter of the data itself than it is of the data's metadata:

- having "FAIR metadata is of very high value in its own right" (FORCE11).

- The data itself might not be accessible, but to find information about the data and its access conditions is already part of FAIRness. (http://blog.ukdataservice.ac.uk/its-not-such-a-fair-way-off/)

- We lack metadata and documentation standards for Social Media Datasets (Hemphill et al. 2019)

Data resources mentioned

- Brena, G., Marco Brambilla, Stefano Ceri, Marco Di Giovanni, Francesco Pierri, Giorgia Ramponi (2019a). News Sharing Users Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions. https://aaai.org/ojs/index.php/ICWSM/article/view/3256/3124.
- Brena, Giovanni; Brambilla, Marco; Ceri, Stefano; Di Giovanni, Marco; Pierri, Francesco; Ramponi, Giorgia, 2019b, "News Sharing User Behaviour on Twitter: A Comprehensive Data Collection of News Articles and Social Interactions", https://doi.org/10.7910/DVN/5XRZLH, Harvard Dataverse, V3.
- Chukwuemeka , David and Abdul, Adeniyi (2017). *The UK 2015 General Election, Twitter data.* [Data Collection]. Colchester, Essex: UK Data Archive. Doi: 10.5255/UKDA-SN-852772
- https://github.com/DataSciencePolimi/NewsAnalyzer
- GE_Readme.docx (http://reshare.ukdataservice.ac.uk/852772/14/GE_Readme.docx)
- 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. https://doi.org/10.18712/nsd-nsd2434-v1
- Ljubešić, N.; Erjavec, T. and Fišer, D., 2017, *Twitter corpus Janes-Tweet 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1142.

References

- Callegaro M., Yang Y. (2018) The Role of Surveys in the Era of "Big Data". In: Vannette D., Krosnick J. (eds) The Palgrave Handbook of Survey Research. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-319-54395-6_23
- Hemphill, L., Hedstrom, M., Leonard, S. (2019) How can we save social media data? Retrieved from http://hdl.handle.net/2027.42/149013
- L'Hours et al (2018) Appraisal/Selection Requirements for New Forms of Data. Deliverable 6.9 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221. https://doi.org/10.5281/zenodo.1406926 . Available at: www.seriss.eu/resources/deliverables
- RDA Fair Maturity Assessment WG. https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
- Japec, L., F. Kreuter, M. Berg, P. P. Biemer, P. Decker, C. Lampe, J. Lane, C. O'Neil, and A. Usher. 2015. "Big Data in Survey Research. AAPOR Task Force Report." *Public Opinion Quarterly* 79(4): 839–80. doi: 10.1093/poq/nfv039.
- Kinder-Kurlanda K, Weller K, Zenk-Möltgen W, et al. (2017) Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data & Society. DOI: 10.1177/2053951717736336
- Kleiner, B., A. Stam and N. Pekari (2015): Big data for the social sciences. FORS Working Papers. https://forscenter.ch/wp-content/uploads/2018/07/fors_wps_2015-02_kleiner.pdf
- Mannheimer, S., and Hull, E.A (2017) Sharing Selves: Developing an Ethical Framework for Curating Social Media Data, International Journal of Digital Curation, 2017, Vol. 12, Iss. 2, 196–209.
- Weller, K., and Kinder-Kurlanda, K. (2016) A Manifesto of Data Sharing in Social Media Research. DOI: http://dx.doi.org/10.1145/2908131.2908172

Questions?

Comments?!

Sugesstions!

Thank you!

University of Ljubljana

Faculty of Social Sciences

Social Science Data Archive

Kardeljeva ploščad 5

1000 Ljubljana

Slovenia

www.adp.fdv.uni-lj.si

arhiv.podatkov@fdv.uni-lj.si

Arhiv.Druzboslovnih.Podatkov

@ArhivPodatkov