gesis

Leibniz Institute
for the Social Sciences

A Manifesto for Data Sharing in Social Media Research

Katrin Weller
Katharina Kinder-Kurlanda

ACM Web Science 2016

Member of the Leibniz Association

WHY?

1. Advance reproducibility and comparability

2. Avoid ‘digital divides’ in data accessibility

3. Save time and money in data collection

HOW?

Successful data sharing

Depends on three critical perspectives:

- Methodological:

  documentation to ensure reproducibility

- Legal:

  share in accordance with various legal frameworks, e.g. data protection legislation, copyright

- Ethical:

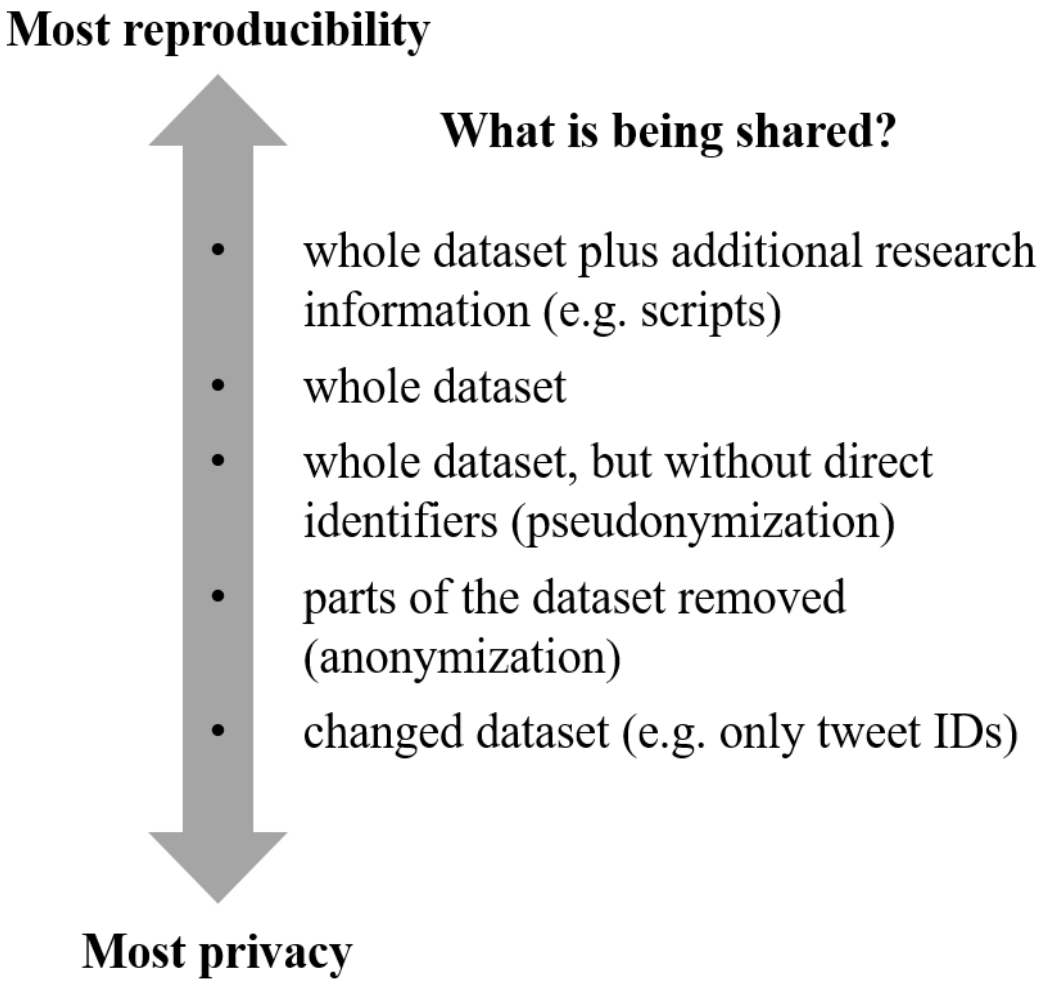
  consider privacy expectations, sharing should „do no harm“

Examples of current practices

Approach	Example	Retrievability	Documentation + standards	Long-term perspective
„Grey market“	Data shared with colleagues (often upon request)	**Low**, via personal connections	**Low**, no standardized documentation or data format	**Low**, no guaranteed long term availability
Researchers‘ personal/ professional websites		**Medium**, URLs may be referenced	**Low**, no standardized documentation or data format	**Low**, no guaranteed long term availability
Social media providers	e.g. Wikipedia dumps	**High**, if directly provided from social media platform	**Medium/high**, depending on the producer	**Medium/high**, depending on the producer
Project-based or thematic collections	e.g. KONECT, CrisisLex	**Medium**, URLs may be referenced	**Medium**, if same principles are applied within entire collection	**Low**, no guaranteed long term availability
Conferences & journals	e.g. ICWSM datasets with conference papers	**High**, usually related to accepted publications	**Medium/high**, depending on the publisher	**Medium/high**, depending on the publisher
Professional archives	e.g. datasets at GESIS data archive	**Medium/high**, datasets may be referenceable with DOIs.	**Medium/high**, depending on the archive’s requirements	**High**, guaranteed availability for different time spans

- For additional examples see our paper: http://dx.doi.org/10.1145/2908131.2908172
- See also: Thomson, S.D. 2016. Preserving Social Media. DPC Technology Watch Report. Retrieved from http://dpconline.org/publications/technology-watch-reports

How much should I share?



WHAT'S NEXT?

Open challenges & next actions

Stakeholder	Open challenges	Next actions
Researchers (individuals/ groups)	• Better documentation practices of how the data was collected *exactly*	• Share internally, i.e. within groups and with collaborators • Share data (and code), in particular datasets used for accepted papers
Archival institutions	• Practices for anonymization	• Develop standards for documentation and metadata • Establish different models for data access (e.g. embargo, secure data access)
Publishers / Conference organizers	• Ways to include checks for data quality in the review process	• Publish non-standard works, such as descriptions of datasets, reproduction studies etc.
Research associations	• Dialogue with major social media companies	• Continuous dialogue with members on best practices, guidelines

Open challenges & next actions

Stakeholder	Open challenges	Next actions
All	• Raise awareness of legal and ethical issues	• Discuss data sharing practices

gesis Leibniz Institute for the Social Sciences

Questions and Feedback

Dr. Katrin Weller
katrin.weller@gesis.org
@kwelle
http://katrinweller.net

Dr. Katharina Kinder-Kurlanda
Katharina.kinder-kurlanda@gesis.org
@ka_kinder
http://www.gesis.org/sdc