New Data Types in Social Science Research and Data Archives

Libby Bishop, Ph.D.
GESIS–Leibniz-Institute for the Social Sciences
Köln

Strengthening and Widening of the European Infrastructure of Social Science Data Archives

5 November 2019

cessda.eu @CESSDA_Data
The challenge of new forms of data

- How can data repositories handle, or prepare to handle, new forms of digital social data?
  - Internet data, social media, web trace, tracking, geo location
  - Facebook – posts, comments, likes
  - Twitter – Tweet, hashtag, followers
  - A few images, to millions of Tweets...
Four questions

1. What are researchers doing with social media and other data?

2. What are repositories currently doing to hold and share new forms data?

3. Are there useful resources for repository staff, and what next steps are planned?

4. What is at stake? What are our responsibilities in the broader debates?
1. What are researchers doing with social media and other data?

- Diverse topics - consumer behavior, hate speech, health, electoral behavior...
- Ready-made tools (e.g., COSMOS)
- Direct access through API or apps (esp. for Facebook)
- Self-collection & sharing by study participants
- Buy from the company or data resellers

*all of these methods differ in transparency and replicability and produce different data and metadata*
Linking and Scraping


“The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships” Breuer, Kinder-Kurlanda, and Bishop (New Media and Society, forthcoming)
2. What are repositories currently doing to hold and share new forms data?
Good reasons for caution

- **Legal** (e.g., copyright, terms of service, ownership of data)
- **Ethical** (e.g., privacy, anonymization, informed consent, 3rd parties)
- **Practical** (e.g., storage space, updates, quality checks)
  - **Volume**: Large amounts of data
  - **Variety**: Different data types and formats
  - **Velocity**: (New) Data generated with high speed
  - **Veracity**: Trustworthiness/quality of data?
- **Documentation** (e.g., (in)compatibility of metadata with existing standards)
Moving carefully....

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

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
Shared FAIRly

“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 datorium (self-archiving GESIS)
  - Findable – Pfeffer, J. and Morstatter, F. (2016)
  - Preserved – DOI - (http://dx.doi.org/10.7802/1166)
  - Reproducible - Python scripts, tools, and documentation
- As open as possible, closed when necessary

As open as possible, closed when necessary
Other more liberal views...

“At George Washington (GW) University Libraries, we (unofficially) interpreted this to allow sharing Twitter datasets that we collected with anyone affiliated with GW (including students, faculty, and other researchers) and their collaborators.”

3. Are there useful resources for repository staff, and what next steps are planned?
Small, Heather, et al. “What your tweets tell us about you” IJDC 2012

Staff at UCLA Library used Association of Internet Research Guidelines to assess risks of curating a Twitter dataset

Hypercities Egypt – Twitter based content documenting the Arab Spring in Egypt and Libya in 2011.

Guidelines on the use of social media data in survey research
CESSDA Work Plan 2020–New Data Types

- A list with relevant elements for documentation and **metadata fields** for social media data. For example, data collection method (if API, what version, search parameters).
- Design of a **plug-in for COSMOS** (researcher tool for accessing Twitter data) that creates an “archiving package”
- **IASSIST session** on the archiving of social media data
- A **paper** on best practices for obtaining informed consent for linking survey data with social media data, including GDPR
- “**How to**” guide on linking surveys and digital trace data in the GESIS Survey Guidelines series
- **Training event** on consent, linking, with Training WG

*Czech Republic, Germany, Greece, Hungary, Slovakia, Slovenia!*
Resources

- **CESSDA DM Expert Guide**
  - Next version for SP staff

- **Managing and Sharing**

- **Publishing Twitter Posts**
  - Why “public” is not enough
  - Williams, Burnap and Sloan, *Sociology*, 2017

- **Documenting Georeferenced Social Science Survey Data: Limits of Metadata Standards and Possible Solutions**
  - Jünger, Borschewski, and Zenk-Möltgen (2019)
  - [https://doi.org/10.1080/15420353.2019.1659903](https://doi.org/10.1080/15420353.2019.1659903)
4. What is at stake? What are our responsibilities in the broader debates?
Raj Chetty 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
Thank You and Questions

Dr Libby Bishop
ElizabethLea.Bishop@gesis.org