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

cessda

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

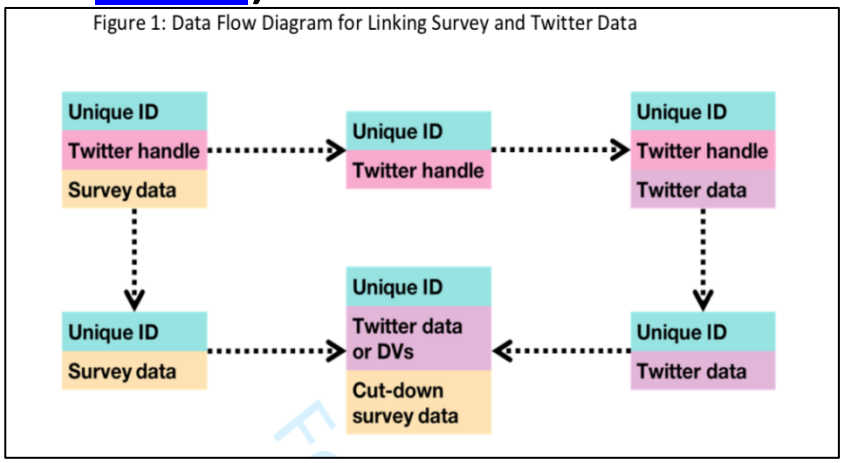
1. What are researchers doing with social media and other data?
2. What are repositories currently doing to hold and share new forms data?
3. Are there useful resources for repository staff, and what next steps are planned?
4. What is at stake? What are our responsibilities in the broader debates?

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

- Linking Twitter accounts to UK longitudinal survey (Understanding Society) Baghal, et al. 2019).



- "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)*

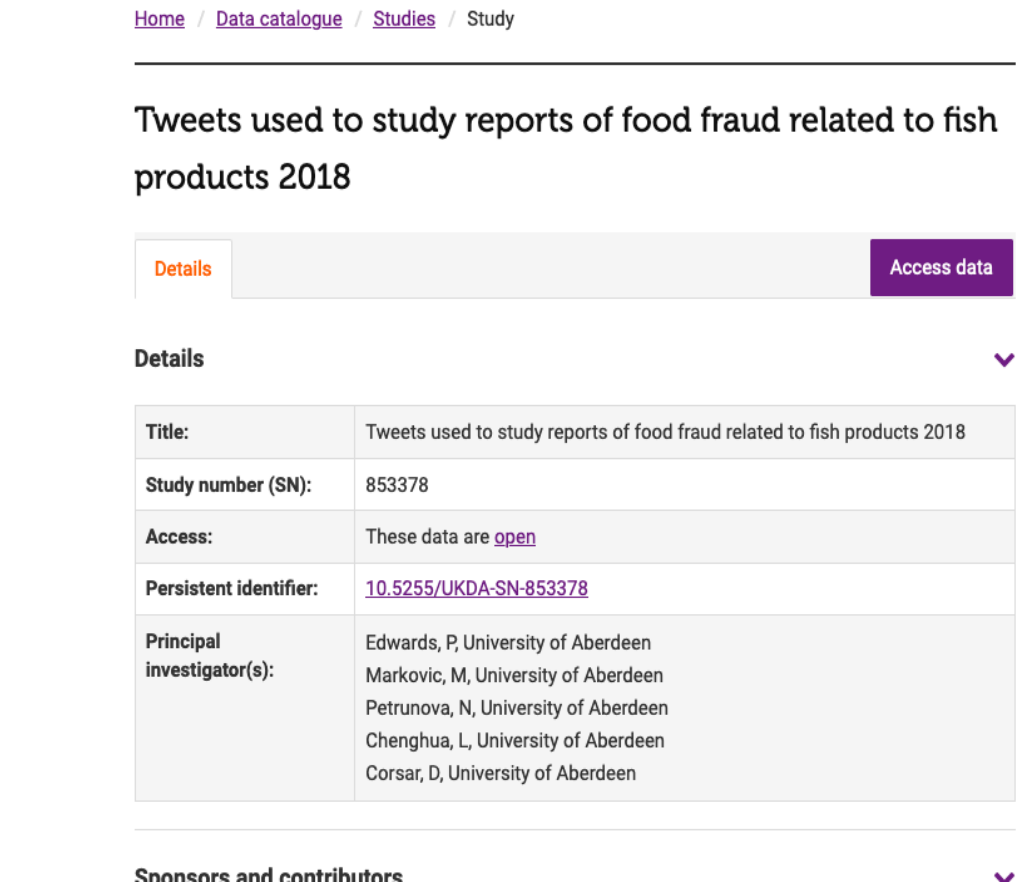
2. What are repositories currently doing to hold and share new forms data?

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….

```
_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
366722915909124098,455624784
366725431057055744,259943580
366725488221237248,17201687
366725912256970752,455624784
366726144768229376,455624784
366727656085004288,265379189
366728312774606850,378242224
366730103427842049,98512503
366731462453641216,184178804
366732788776439810,49618309
366733754980503554,64516663
366734911039750146,98512503
```



Home / Data catalogue / Studies / Study

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

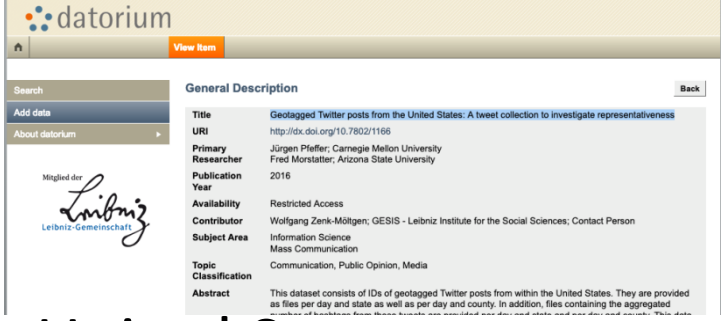
Details | Access data

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

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

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.”

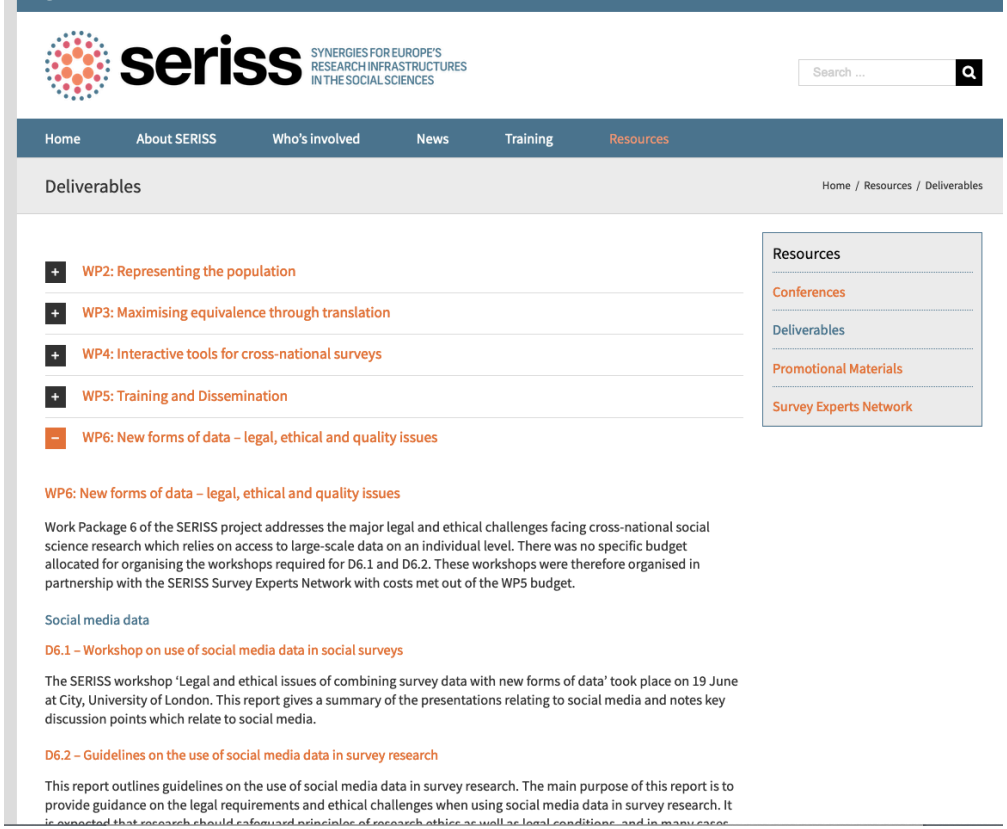
- (Justin Littman, “Twitter’s Developer Policies for Researchers, Archivists, and Librarians” https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2)

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

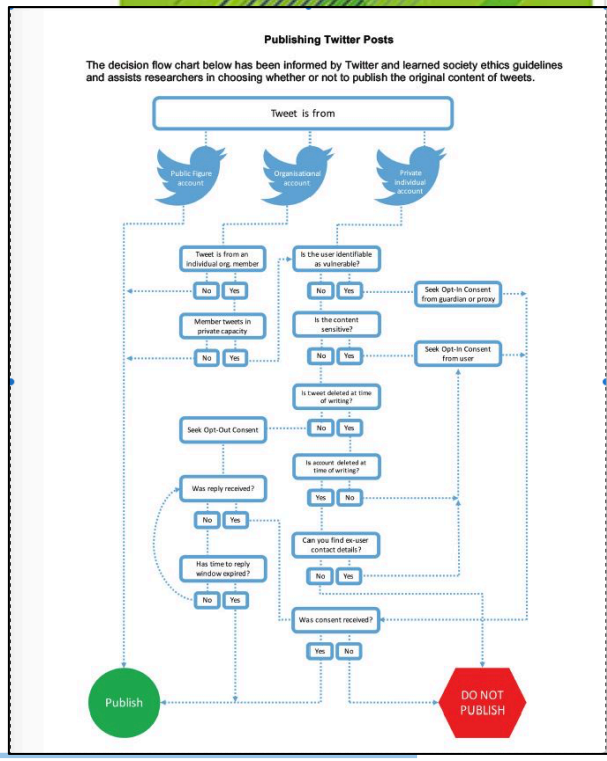
Czech Republic, Germany, Greece, Hungary, Slovakia, Slovenia!

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



4. What is at stake? What are our responsibilities in the broader debates?

Should we do more to make data FAIR?

- 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

cessda.eu @CESSDA_Data

cessda