



DIGISURVOR Workshop 1

‘Linking Digital Footprint and Survey Data for Open Research’

**UNIVERSITY OF MANCHESTER, 13-14TH FEBRUARY 2025.
BOARD ROOM, 2ND FLOOR, ARTHUR LEWIS BUILDING**

13.30 – 15.15	Session 1: Introductions and the DIGISURVOR project – an Overview UoM team - Rachel, Marta, Alex, Riza and Conor
15.15 – 15.30	Coffee break
15.30 – 17.00	Session 2: Overview of existing data donation platforms Bella Struminskaya (Utrecht University) – The EU - The D3I Project Alexi Quintana Mathé (Northeastern University) – The US - National Internet Observatory David Zendle (University of York) – The UK - Data Donation Service (DDS) UK
From 19.00	Workshop dinner – Gusto
09.00-09.10	Arrivals and coffee
09.10 – 10:45	Session 3: Linking survey and social media data Jonathan Nagler (NYU Center for Social Media and Politics) Tarek Al Baghal (University of Essex) Andreu Casas (Royal Holloway) Sebastian Stier (GESIS)
10.45 – 11.00	Coffee break
11.00 – 12.30	Session 4: Linking survey and web-tracking data Melanie Revilla (Universitat Pompeu Fabra) Florian Keusch (University of Mannheim) Toby Crisp (IPSOS) Kiran Arabaghatta Basavaraj (UCL)
12.30 – 13.15	Lunch
13.15 – 14.45	Session 5: Social science research using linked data Reuben Bach (GESIS) Silvia Majo-Vazquez (Vrije Universiteit Amsterdam) & María Victoria-Mas (Universitat Abat Oliba CEU Barcelona) Sarah Shugars (Rutgers University)
14.45 – 15.00	Coffee break
15.00 – 16.00	Session 6: Group Discussion
16.00 – 16.15	Close

Introducing DIGISURVOR: “Linking Digital Footprint and Survey Data for Open Research”

- The main goal of DIGISURVOR is to investigate the feasibility of producing datasets for open research* that integrate individual-level survey data with DTD.
- We do so using three existing datasets that combine survey data with two types of individual-level DTD – social media feed content (2 US datasets) and (domain level) web URLs (1 UK dataset).
- From these data, we will generate range of new observational variables based on respondents’ digital transactions that will augment, enhance, and help to validate their survey responses.
- Specifically we focus on conceptualising, operationalising and constructing a set of ‘core’ DTD-based variables that can be generated from individuals’ social-media and/or web-browser data and linked to their survey responses that maintain respondent anonymity.

*We distinguish here between fully "open data" which is data that is freely available and anyone can access, use or share, and data that supports open research, i.e. that increases the public value of these types of datasets, by making them more findable, accessible, interoperable and reuseable (FAIR).

The Datasets: Overview

The three datasets we use covering the period 2020-2024 and link individual survey responses and their DTD. All were collected for analysis of substantive research questions, as part of independent externally funded projects i.e. not the specific methodological questions posed in this project.

Type 1: Linkage of respondents survey and Twitter/X data. (2 – US respondents)

Type 2: Linkage of respondents survey and web browsing data (1 – UK respondents)

Dataset: Type 1

We use two datasets of this type collected in the US as part of a European Research Council funded project - DiCED. They both combine a two-wave pre and post-election panel survey responses with individuals Twitter data (now X). They were collected at two time points - 2020 (dataset 1a) and 2024 (dataset 1b).

The survey component measures media consumption, perceptions of digital campaign contact, awareness of misinformation, core political attitudes and behaviours, plus standard socio-demographic characteristics and Twitter use.

Dataset 1a will be used as a training dataset to identify and extract two subsets of anonymised and standardised data from respondents' tweets and Twitter accounts that can be augmented to the survey data. The results of that exercise will be re-run with dataset 1b for purposes of replication and validation.

Dataset 1a US 2020

2-Wave YouGov panel study. Wave 1 was fielded to over 5,000 respondents between 16 September – 20 October 2020, and wave 2 was fielded in the week following the election (9 November). Respondents to the pre-election survey were invited to share their Twitter handle with the research team. 1,598 respondents agreed in the pre-election (wave 1) to share (27% of the overall sample). There were two main stages of attrition: 361 either did not subsequently provide a handle or provided an empty /incorrect handle. Of the remaining 1,237, 920 could be validated against the Twitter API, which constituted 15% of the overall sample. Of these 920 accounts, 697 individuals completed the post-election survey.

We collected the tweets, retweets, follows and likes of these respondents during the period 1 September 2020 – 9 November 2020. We also collected the timelines of the accounts that respondents were following during this period to construct a measure of exposure to presidential election campaign content on Twitter.

Dataset 1b US 2024

2 Wave YouGov panel study. Wave 1 was fielded to over 5,000 respondents between 24 September – 21st October 2024, and wave 2 was fielded post-election (12-26 November). Respondents to the pre-election survey were invited to share their Twitter handle with the research team. 1306 respondents agreed in the pre-election (wave 1) to share (23% of the overall sample). There were again 2 main stages of attrition: 127 did not subsequently provide a handle/provided an incorrect or empty handle. Of the remaining 1,179, 964 could be validated against the Twitter API, which constituted 15% of the overall sample. Of these 964 XX accounts, individuals completed the post-election survey.

We collected the tweets, retweets and follows of these respondents during the period 24 September 2020 – 26 November 2020. Due to the new cost structure on X data collection post Musk, we were not able to collect the timelines of the accounts that respondents were following.

Dataset: Type 2

- Web-tracking data:

- Collected by YouGov in the UK
- Collected from mobile devices between 20 March-21 May 2020 (9 weeks)
- Includes URLs visited by participants that have been classified as “news navigation”
- URLs only available at domain level
- And information about referral apps: social media sites, messaging apps or Google.

- Linked to a two-wave panel survey:

- Wave 1 (N=597) - after 5 weeks of tracking, measuring political attitudes, media habits and trust, demographics.
- Wave 2 (N=499) - final day of tracking, repeated questions about political attitudes, media habits and trust.

Overall, this dataset provides a unique opportunity to link self-reported political attitudes and perceptions of the media to observed measures of online news consumption.

Analysis

Analysis of the datasets will occur in two phases.

Phase (1) Proof of concept: we will design and generate a range of new attitudinal and behavioural variables from individuals' DTD that maintain respondent anonymity and that can be used to a) validate and b) augment and enhance the survey responses.

Phase (2) Proof of value: we will investigate the newly generated DTD variables for sources of bias, i.e. device coverage, response rates and measurement error.

Phase (1) Proof of concept – Datasets 1a and 1b

Subset 1: Structural variables

A range of 'core' anonymised variables from individuals' accounts (tweets, meta-data) to describe the content and structure of their tweets e.g. average length (characters, words), use of hashtags, mentions, URLs, acronyms and abbreviations, and emojis. The frequency and mode of their activity e.g. authoring vs retweeting; length of membership, number of posts, number and ids of followers and accounts followed.

Subset 2: Substantive variables

A further set of substantive variables measuring respondents attitudes and behaviours will be generated from the DTD for purposes of:

- **2a Methodological validation**
- **2b Investigation of subject-specific research questions.**

Phase (1) Proof of concept – Dataset 2

Variables measuring the characteristics of individuals' news consumption in **3 key dimensions**:

A) Volume and frequency of news consumption – e.g. number of visits and time spent on news sites.

B) Fragmentation and ideological diversity of the news diet – e.g. Simpson's D, Shannon's H indexes, Partisan Skew score etc.

C) Credibility of websites visited – Harmonisation of existing credibility scores for news websites (or scores produced if not available). Aggregate measure of overall credibility score of individual's news repertoire.

Where feasible, we will ensure that the methodology used to develop variables is **transferable** to datasets 1a and 1b (e.g. index of credibility of the news shared by individuals through their Twitter feeds).

What we have done so far

Variable Construction (Dataset 1a/1b)

So far, a preliminary list of variables has been drawn up for Dataset 1a/1b (Twitter/X data) which is divided into three categories:

1. **Structural:** A set of core anonymous variables that can broadly *describe* the general content and structure of a respondent's digital data.
 - This includes general profile metrics such as **length of membership**, **number of followers**, **posting frequencies**, **post category types**, **use of attachments**, and **average post engagement**.
 - It also includes more complex variables such as **post topics**, **profile formality**, **posting toxicity**, **posting sentimentality**, **network credibility** and **influencer status**
2. **Substantive:** A more complex set of additional variables that measure/estimate key political concepts, attitudes or behaviours.
 - These involve estimating new variables of interest based on key digital indicators. For example, **ideological position**, **populist position**, **candidate/policy endorsement**, **political attention**, and **misinformation exposure**.
3. **Validation:** A set of variables that exist in both the survey data and digital data in some form, where one can be used to validate the other.
 - This includes validating responses to **media consumption** and **ideological diversity**, as well as claims about **online usage**, perceptions of **misinformation exposure** and **filter bubbles**, and **self-reported ideological position**

Example Variables (Dataset 1a/1b) – 56



STRUCTURAL (22)	SUBSTANTIVE (17)	VALIDATION (17)
<p>Post Topics: What do users typically post about? Is it primarily political, sport, music, film, work, general life? etc. What do they post about <i>within</i> topics? (e.g: within politics)</p> <p>Post Categorisation: What type of general user are they? Are they primarily a “solo author” (mostly original tweets), an “amplifier” (more RTs), or an engager (Replies/QTs/Mentions)? Are they a “lurker” (mostly likes only)?</p> <p>Profile/Post Formality: How “formal” is a user’s profile and posting behaviour? Do they mostly use informal language, poor grammar and a high % of abbreviations, hashtags and emojis?</p>	<p>Left-Right Ideological Position: Can we generate measures of ideological position based on a user’s digital data? Based on follower networks, content engagement, or content they post about.</p> <p>Populist Sentiment: Can we generate estimates of degree of populism based on the content in a user’s posts? Methods for measuring anti-establishment rhetoric, etc.</p> <p>Content/Policy Endorsement: We can directly quantify a user’s position on particular policies or individuals using methods for sentiment or stance classification.</p>	<p>Privacy Paradox: How private or anonymised is a user’s digital profile? How much sensitive information do they give away and how does this marry up to their reported concerns about data privacy?</p> <p>Filter Bubble Perception: What is the ideological diversity of the users they follow and the information they are exposed to? How does this marry up with their reported ideological news diet?</p> <p>Online Network Quality: How “credible” is the network of accounts a user follows and the information they post? How does this marry up with a user’s perception of online network quality?</p>

How AI/NLP can help in estimating variable values

Substantive Variables that assess the extent to which content is **political**

- Online Political Attention/Discussion
- Political Campaign Exposure/Engagement

Potential supporting tools: classification (or regression) models trained to identify political posts (or posts pertaining to political campaigns)

- would require a corpus (text collection) where text examples are labelled as *political/not political* [1, 2]
- likely to not be a very difficult classification/regression problem (i.e., an objective task)
- complexity: 
- reliability: 

[1] The Twitter Political Corpus. Available from: <https://www.usna.edu/Users/cs/nchamber/data/twitter/>

[2] The PoliBERTweet Dataset: Available from: <https://github.com/GU-DataLab/PoliBERTweet>



Icons by Najmun Nahar (Freepik)

How AI/NLP can help in estimating variable values

Substantive Variables that detect **ideological leaning**

- Content/Person Endorsement (stance detection)
- Ideological Position/Exposure (political scaling)
- Populism Propagation/Exposure

Potential supporting tools: classification models trained on a pre-defined typology

- would require a training corpus [3-5]
- stance detection can suffer from lack of context, e.g., sarcasm, irony (i.e., can be a subjective task)
- political scaling can be challenging (depending on how many classes there are in the typology)
- complexity: 
- reliability: 

[3] The P-Stance Dataset. Available from: <https://github.com/chuchun8/pstance>

[4] The Senator Tweets Dataset. Available from: <https://huggingface.co/datasets/m-newhauser/senator-tweets>

[5] The Us vs. Them Dataset: Available from: <https://github.com/LittlePea13/UsVsThem>



Icons by Najmun Nahar (Freepik)

How AI/NLP can help in estimating variable values

Substantive Variables that detect any **harmful content**

- Misinformation/Disinformation Propagation/Exposure
- Conspiracy Theory Propagation/Exposure

Potential supporting tools: classification models trained on a fact verification task

- would require a training corpus [6, 7] and a knowledge base with factual information (e.g., Wikipedia, mainstream news agencies)
- complexity: 
- reliability: 

[6] The MiDe22 Dataset. Available from: <https://github.com/metunlp/MiDe22>

[7] The TruthSeeker Dataset: Available from: <https://www.unb.ca/cic/datasets/truthseeker-2023.html>

Icons by Najmun Nahar (Freepik)

Next steps – Proof of value phase

Understanding data quality

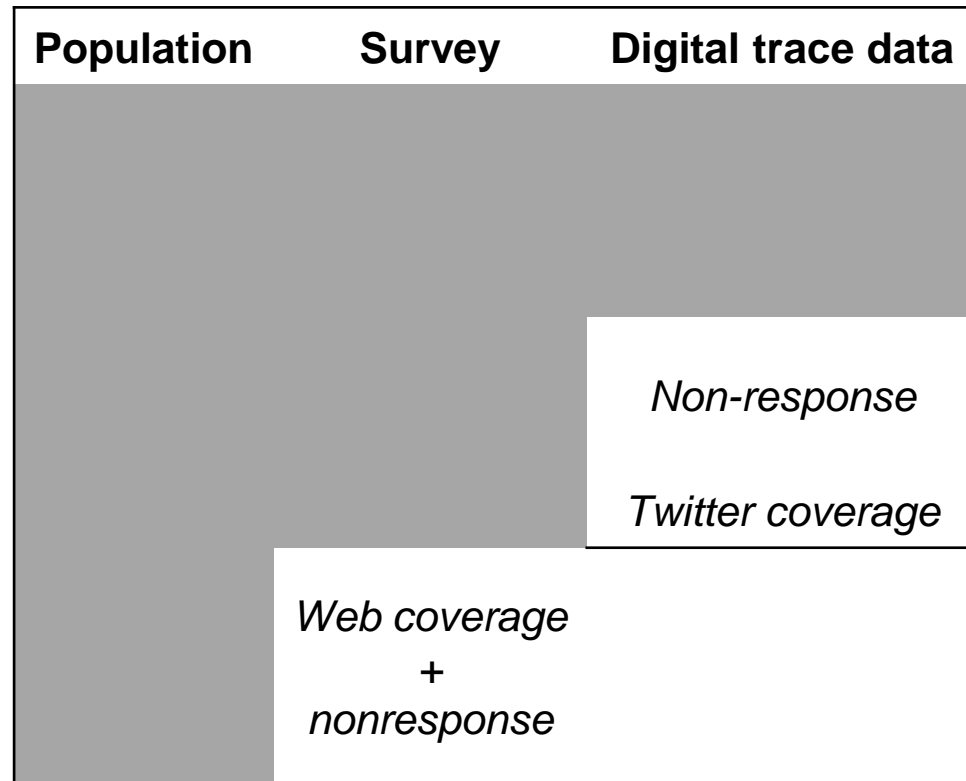
Selection bias

- **Survey**: coverage error, sampling error, non-response error
- **Digital trace data**: DT coverage error, DT non-response error

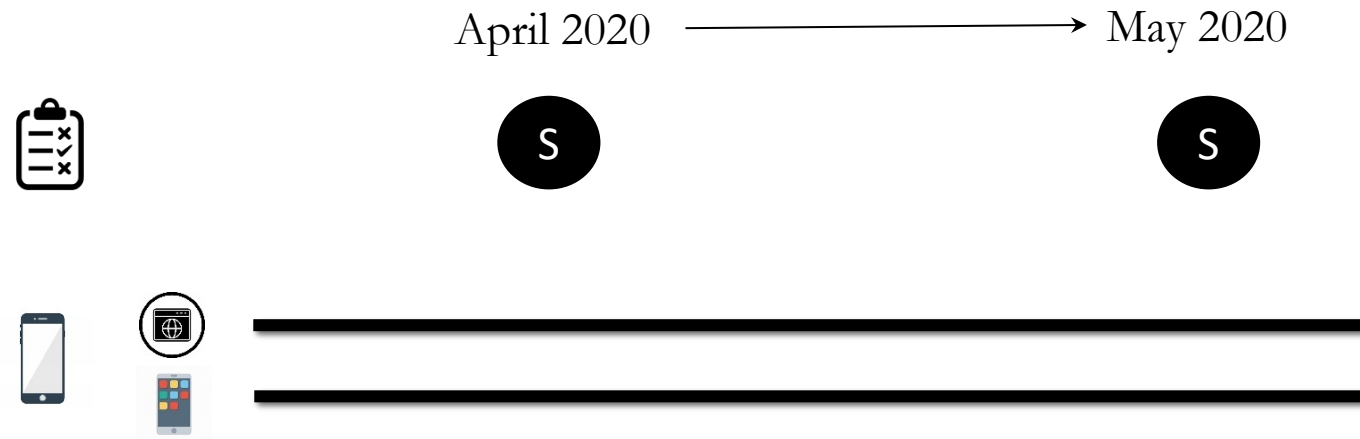
Measurement bias

- **Survey**: validity, measurement
- **Digital trace data**: DT validity, DT measurement

Identifying non-response in digital trace data



Identifying measurement error



Methods for estimation and correction for error

Selection error:

- Comparison with population estimates, regression models
- Non-response weighting and raking

Measurement error:

- MultiTrait-MultiMethod
- Saving plausible values of latent measures

Questions for discussion groups

1. Do our variables capture the main characteristics of interest regarding individuals' Twitter use?
2. Does our traffic light/coding system work – is it a sensible approach to identifying the most important variables to add to the survey data - MIV
3. What other variables might be included that we have currently not identified?
4. How 'generic' are these variables in terms of cross-platform application. Could they be transferred to other new forms of DTD – i.e. Facebook, Instagram, YouTube, Reddit, TikTok?

You can get a copy of this presentation from:

<https://tinyurl.com/digisurvivor-pres>

You can use the URL corresponding to your group below, to write down your answers:

<https://tinyurl.com/digisurvivor1>

<https://tinyurl.com/digisurvivor2>

<https://tinyurl.com/digisurvivor3>

<https://tinyurl.com/digisurvivor4>

Summary & Next steps

Different models of DTD and survey data linkage collection and access

- Decentralised /distributed projects collecting linked DTD and survey data. E.g DiCED
- Data donation resource centres – software provision / open source to enable and support data linkage. E.g D31
- Centralised research/non profit providers at the national level being established based on some model of DD and some system of user accreditation – secure environment - to allow access and analysis of data. Export permitted only at aggregate level. E.g. NIO, SDDS
- Platform provided option SOMAR at ICSPR/Michigan – again vetting process and ‘clean room’ built to allow for approved researchers to use the data.
- Commercially provided option panels – proprietary /paid for access. Managed in house, ethical compliance systems internal e.g. IPSOS

Open research data challenge

Distributed model – typically not focused /resourced to enable open dataset production.

National non-profit service / platforms/ commercial providers – secure rooms get round this to an extent but still operate limited system of access, privileged access for approved researchers'

Still not ideal – Journals require datasets to be deposited for re-analysis; Datasets using public funds are required to be deposited at National Archives, UKDS. E.g. national and international infrastructure surveys – BES, US, ESS if they develop a linkage component how do we equip them to be shared if they include linked data. Only through safe rooms?

Still a need a model for extracting variables from DTD that value add that can be made 'public' or at least shareable with a broader range of users working outside a safepod or controlled environment. Need criteria to measure the robustness of that data – bias detection and correction.

This workshop helpful to locating DIGISURVOR project in that space.... Currently end point in the cycle and largely an after thought rather than a primary concern.

Should be built into or baked into the linkage process. What types of standardized and anonymized data should data producers generate as part of their project so that they can share it?

Tiered or hierarchy of access. Inner core data owners/controllers; Approved users in controlled environment; Approved users in uncontrolled environment; Fully open data

Follow ups

Reflect on the workshop presentations and discussion and implications for DIGISURVOR e.g. SNA, personality measures, cross-platform application and over time and space generalization 'future proofing';

Github repository – share documents /code/outputs e.g. MIV variable list coded structural/substantive/validating, and by complexity, reliability and usability; conference paper;

Send the url to all workshop participants – create an email list. Let us know if you want to opt to sharing your email address.

Workshop No.2 12 months time – report back on our progress and your progress!

Building a network of interested researchers

