DIGISURVOR Workshop 2 ‘Linking Digital Footprint and Survey Data for Open Research’

UNIVERSITY OF MANCHESTER, 13-14th January 2026.

12.30 – 13.15	Arrival Lunch
13.15 – 13.30	Welcome and introductions
13.30 – 14.00	**Session 1: Update on the DIGISURVOR project** UoM team - Rachel, Marta, Alex, Riza and Conor
14.00 – 15.00	**Session 2**: **Keynote talk** "*Linking Survey and Social Media Data: Bridging the Gap Between Data Protection and Open Research"* Presentation UoM team and group exercise
15.00 - 15.15	Coffee break
15.15 – 16.30	**Session 3: Detecting and correcting bias in linked data sources** **Chair: Marta Cantijoch** Paulina Pankowska (University of Utrecht) and Ruben Bach (University of Mannheim) *"The gendered division of cognitive household labor and mental load in the digital space"* Sarah Shugars (Rutgers University) **"***The speech we miss: How keyword-based data collection obscures youth participation in online political discourse"* Conor Gaughan and Alex Cernat (University of Manchester) "*Who consents to sharing their tweets with researchers? A comparative analysis of selection bias in linked survey and social media data*."
16.30 - 17.30	**Session 4: Roundtable Discussion: Researcher Access to DTD - Future Prospects, Opportunities and Challenges** **Chair: Rachel Gibson** Kate Dommett (University of Sheffield) Co-Chair UKRi Social Platforms Data Access Taskforce Veronica Guzman Quilaqueo (UKRI, Senior Strategy and Partnerships Manager SDRUK) Andreu Casas (Royal Holloway, University of London) Director, London Social Media Observatory
17.30 – 19.00	Free time / hotel check ins
From 19.00	**Workshop dinner –** Tai Wu | Manchester Chinese Restaurant 81-97 Upper Brook St

09.30 – 11.00	**Session 1: Using Linked DTD News Consumption** **Chair: Alex Cernat** Andreu Casas et al. (Royal Holloway, University of London) *"Comparing Misinformation Inoculation Interventions: Fact Checking, Media Literacy, and High-Quality News Boost"* Sílvia Majó-Vázquez et al. (Vrije Universiteit Amsterdam) *"Measuring Online News Audience Fragmentation and Ideological Segregation Across Countries Time and Media Systems"* Marta Cantijoch and Conor Gaughan (University of Manchester) "*Content-Based Classification of URL Domains by Large Language Models*" Jonathan Nagler (NYU) *"Simple Aggregates from Digital Trace Data Donations to Merge with Survey Data and Examine Cross-Platform Media Consumption"*
11.00 – 11.15	Coffee break
11.15 – 12.15	**Session 2: Designing Software & Tools to collect and analyse DTD** **Chair: Conor Gaughan** Diana Maynard (University of Sheffield) "*Visualising Toxicity: Interactive Dashboards for Social Media Abuse Monitoring*" Riza Batista-Navarro and Thomas Flavel: "*Social Media Mining in KNIME: Democratising Access to Libraries for Text Analysis (DELTA)*"
12.15– 13.15	Lunch
13.15 – 14.30	**Session 3: Developments in Infrastructure to support DTD donation and linkage** **Chair: Riza Batista-Navarro** David Zendle and Faye Chivers (University of York) "*The Smart Data Donation Service: Year 1 of a New Piece of National Research Infrastructure"* Bella Struminskaya (Utrecht University) *"Building sustainable software-centered research infrastructures to support digital data collection"* Steve McEachern and John Sanderson (UK Data Service) "*Digital trace data management in the UK - legal, technical and practical considerations"*
14.30-.14.45	Coffee break
14.45 – 15.30	**Session 4: Practitioners** **Chair: Rachel Gibson** Toby Crisp (Ipsos) "*New Developments in the IRIS panel*" Adam McDonnell Abigail Axel-Browne (YouGov) *"A New Capability - YouGov Behavioural"*
15.30 – 15.45	Closing Comments

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 (Twitter/X) feed content (2 US datasets) and (domain level) web URLs (1 UK dataset).
- From these data, we generate a range of new observational variables based on respondents' digital transactions to augment, enhance, and 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 cover the period 2020-2024 and link individual survey responses and their DTD. All surveys were fielded by YouGov 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 respondent survey and Twitter/X data - collected in the US as part of a European Research Council funded project - DiCED. DTD collected at two time points - 2020 (1 Sept - 9 Nov N = 920 pre and 697 pre/post), Dataset 1a; and 2024 Dataset 1b (24 Sept – 26 Nov N = 964 pre and pre/post). Twitter data collected by research team - tweets, retweets, follows and likes and timelines* The surveys measure media consumption, perceptions of digital campaign contact, awareness of misinformation, core political attitudes and behaviours, plus standard socio-demographic characteristics and Twitter use.

Type 2: Linkage of respondent survey and web browsing data - collected in the UK by Spanish research team led by Maria and YouGov. Collected via mobile devices between 20 March-21 May 2020 (9 weeks). Includes URLs visited by participants classified (post data collection) as “news navigation”. URLs at the domain level, plus 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.

*likes and timelines(2020 only)

Analysis

Analysis separated into two phases.

Phase (1) Proof of concept: 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: 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 Type 1 Linked Survey and Twitter/X data

- **Structural variables -** We identified 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.

- **Substantive variables -** A further set of substantive variables measuring respondent attitudes and behaviours will be generated from the DTD for purposes of Methodological validation and Investigation of subject-specific research questions.

Dataset Type 2 Linked Survey and Web browser data

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

- **Volume and frequency of news consumption** – e.g. number of visits and time spent on news sites.
- **Fragmentation and ideological diversity of the news diet** – e.g. Simpson's D, Shannon's H indexes, Partisan Skew score etc.
- **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).

Dataset Type 1: Linked Survey to Twitter/X Data

- Built a standardized procedure for the extraction of anonymized structural and substantive variables from participant DTD

Variable Level	Overall
Profile-Level	65
Post-Level	61
Overall	126

- To create these variables, we enlisted a range of out-of-the-box computational tools which sought to combine maximum simplicity and replicability (based on our linked Twitter data)

- The aim was to build a transparent and replicable *template* (skeleton code) for others to implement and build upon for their own research

- Code and user guide can be found on the DIGISURVOR webpage: https://digisurvor.github.io/data_and_code/

Dataset Type 2: Linked Survey to Web Browsing Data

- Currently in the process of building an automated procedure for the classification of URL domains by content
- This includes the programmatic use of state-of-the-art LLMs for the out-of-the-box classification of URL domains with zero shot prompting
- Thus far, we have the classification of just over 4,000 common URL domains (UK-centric) into 18 predefined categories
- Currently working to expand on this to include ideological and credibility classification

Phase (2) Proof of value: Progress to date

Understanding data quality

Conor Gaughan and Alex Cernat (University of Manchester) *"Who consents to sharing their tweets with researchers? A comparative analysis of selection bias in linked survey and social media data."*

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

Next steps: Audit of augmented datasets 1 and 2 to identify variables that measure the same concepts both in the survey and the DTD and compare them. Having identified variables, we can build some more advance models to estimate different types of measurement errors.

Digital Footprints Conference	14th May - 15th May	University of Leeds, Leeds, UK
6th Workshop on Mobile Apps and Sensors in Surveys (MASS)	4th June - 5th June	London School of Ecnomics, London, UK
Open Research Conference	9th June - 10th June	Manchester Business School, Manchester, UK
SoSS Research showcase	25th June	University of Manchester, Manchester, UK
11th Conference of the European Survey Research Association (ESRA)	14th July - 18th July	Utrecht University, Utrecht, Netherlands
UK Data Service Seminar	24th July	University of Manchester, Manchester, UK
Royal Statistical Society Conference - SDR UK Panel Discussion	2nd Sept	Edinburgh
Turing Social Data Science WG seminar	12th Sept	Virtual
NLPOR@COLM 2025	10th October	Hybrid - Montreal
Web Science Conference 26	26th-29th May 2026	Braunschweig, Germany
Digital Footprints Conference	19th May - 21st May	University of Leeds, Leeds, UK
NLP + CSS Workshop 2026	6th - 7th July	ACL Conference, San Diego, California

Outputs – Conferences

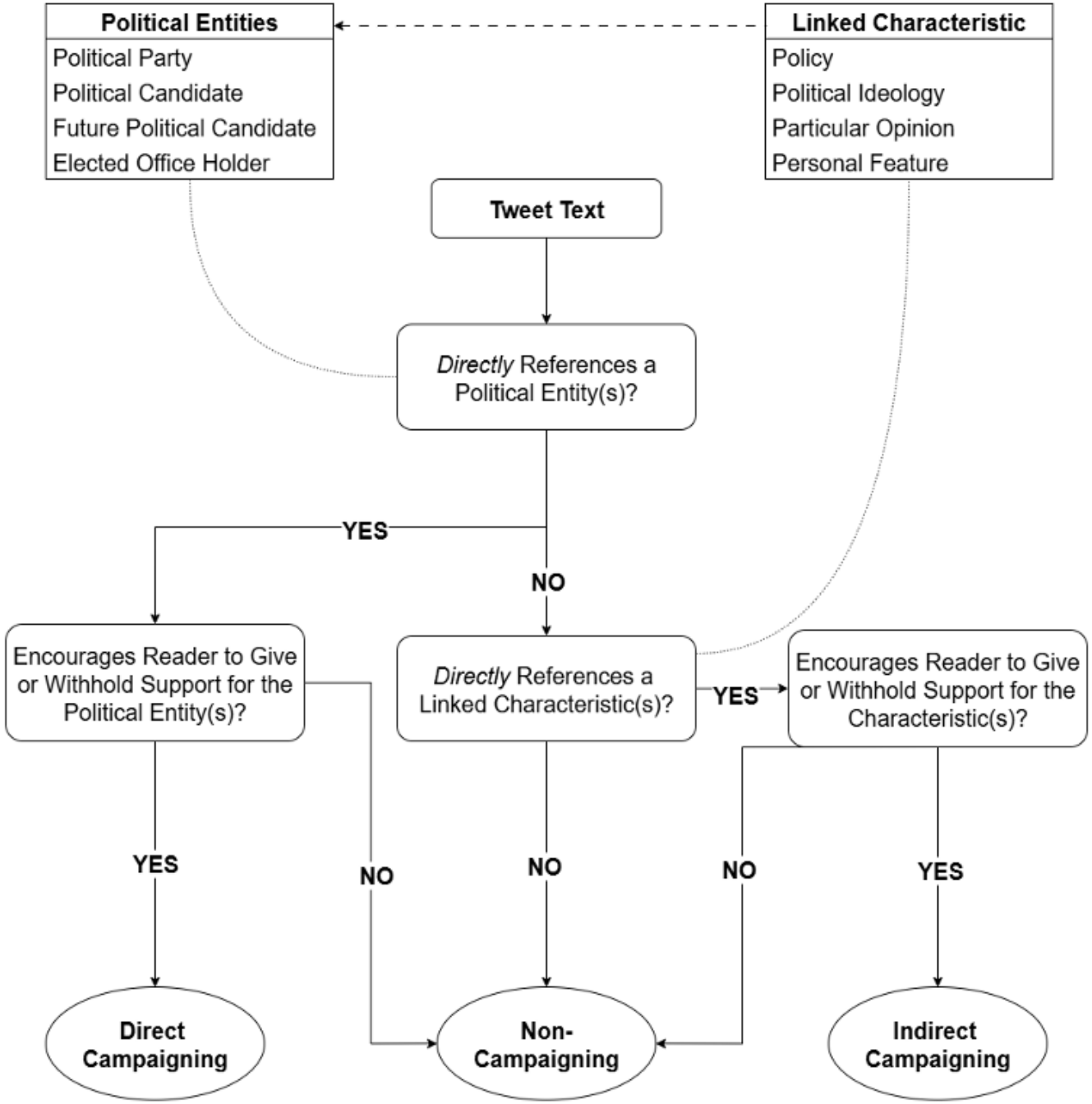
Outputs – Papers

"Linking Survey and Social Media Data: Bridging the Gap Between Data Protection and Open Research." Gaughan et al. (to be submitted)

Researchers working with linked digital trace and survey data typically find themselves in conflict between compliance with data privacy regulations and the values of open research. This paper addresses this challenge by outlining the current approaches to resolving this tension and the trade-offs they entail. Specifically we address two main research questions **RQ1:** To what extent can *standalone* SMD be anonymized while maintaining a degree of utility for other researchers? Additionally, we consider this in the context of linking SMD with survey data and ask a further question: **RQ2:** To what extent can *linked* survey-to-SMD be anonymized while maintaining a degree of utility for other researchers? We generate a new classificatory framework - The Data Access-Data Protection Spectrum - that maps existing and 'ideal' approaches to answer these question.

"Who consents to sharing their tweets with researchers? A comparative analysis of selection bias in linked survey and social media data" (R&R)

Survey research is entering a new era which centres on its linkage with other forms of digitally generated data such as social media. Many suggest that this can help to address existing weaknesses in self-report surveys such as non-response and measurement bias. However, to link a participant's survey responses to their social media data, consent from the participant is required. Previous studies have shown that consent to linkage is typically low and selective. This paper expands on the existing literature by comparing Twitter (now X) usage and consent to survey linkage across five national contexts. Testing the effects of several socio-demographic and attitudinal predictors in the US, the UK, France, Germany, and Poland, our study finds that overall consent rates vary significantly by age, political attention, privacy concern, trust in social media companies, and frequency of political posting on Twitter/X. However, our results also confirm that variable effects differ significantly between nations, suggesting a moderating cultural influence. Within-country variation in the US between 2020 and 2024 is also present, indicating that effects are not necessarily fixed over time. These findings dictate the need for caution when conducting substantive comparisons across countries and time when using social media data.



Ongoing NLP work – Automatic detection of political campaign content

Open research data challenge of working with Linked Survey and DTD

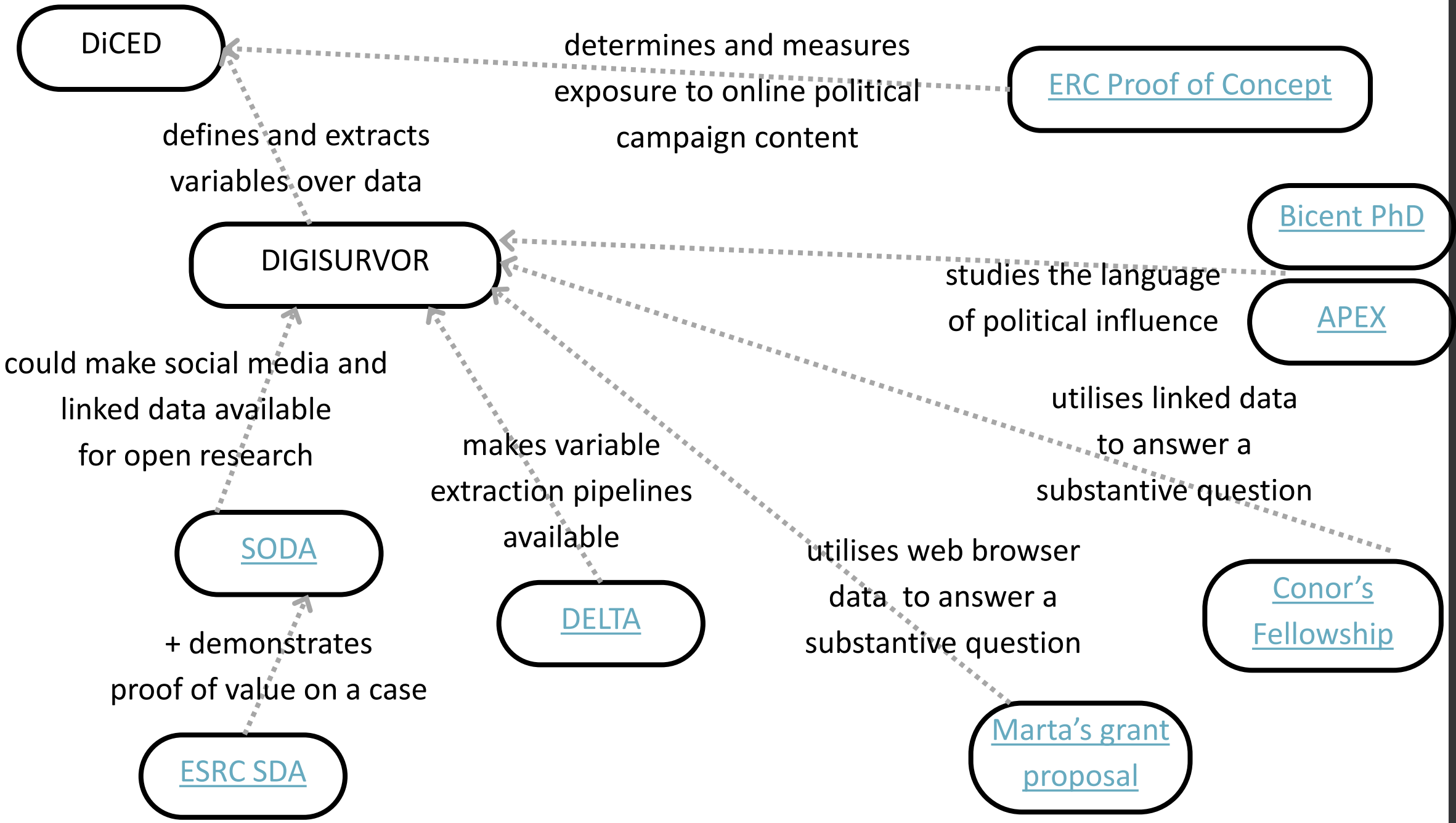
- Ended the first workshop by drawing attention to the tension we identified within DIGISURVOR's goals in terms of the re-use of Linked DTD and Survey data, between data protection and open data principles.

- We realised this tension lay at the heart of the challenge or project so we spent some time reflecting and conceptualising this question. Session 2 will develop our thinking about this.

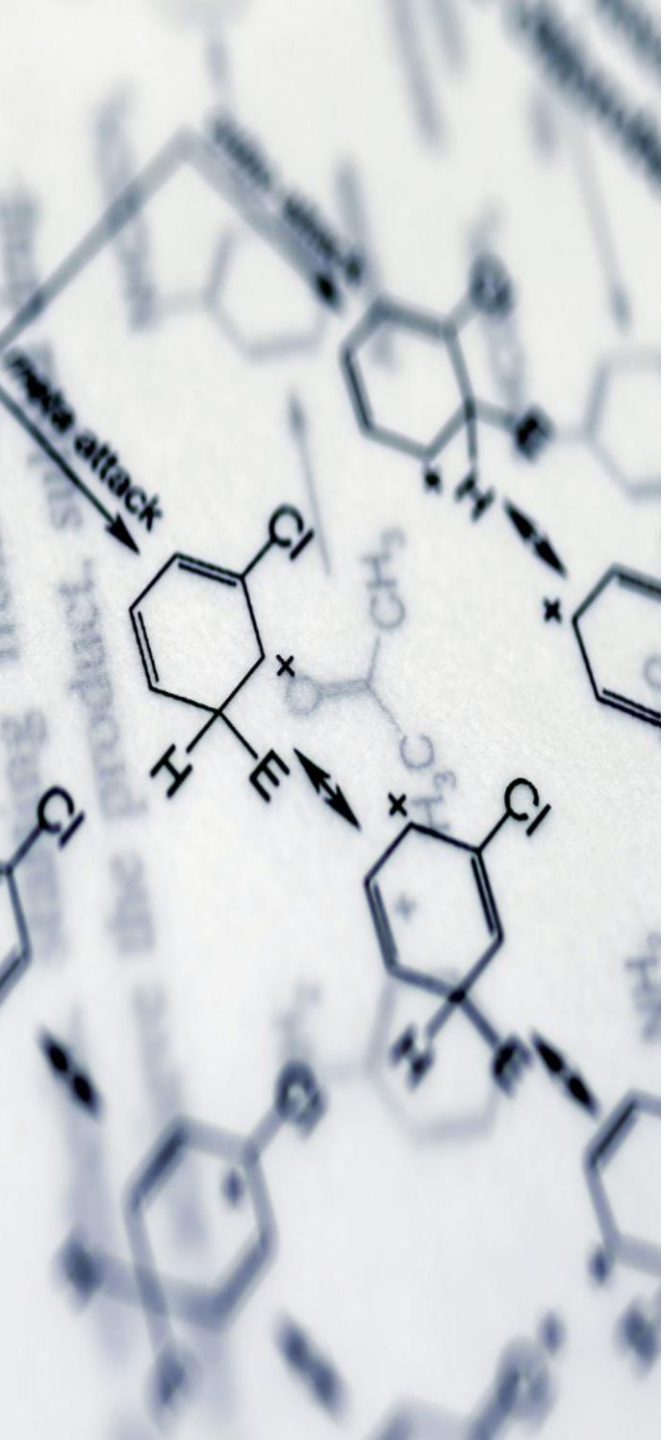
- Also led us down a few avenues we had not anticipated but which were all arguably necessary to navigate in order to properly address the challenges/core questions we set out at the start.

- We began by asking what types of standardized and anonymized data can data producers generate from linked Survey and DTD such that they can share it?

- Subquestions
  - Even if it is possible to produce standardized and anonymized data and share the code underlying the variable construction, how can we make the new augmented datasets available for replication and validation.
  - Could we construct synthetic data from our social media data to allow for replication and would this meet the threshold for anonymisation and motivated intruder testing. What does that threshold look like? (SODA)
  - Is it possible to build an end to end software pipeline for researchers to use that allows them to more easily collect, analyse and share linked survey and DTD (Delta)
  - What types of new substantive questions can we address using linked survey and DTD – The Rhetorical Frames of Online Influencers (PRISM), Exposure to Populist Campaign Content, Measuring Full Exposure to News.


DiCED
determines and measures exposure to online political campaign content
ERC Proof of Concept
defines and extracts variables over data
DIGISURVOR
Bicent PhD
studies the language of political influence
APEX
could make social media and linked data available for open research
makes variable extraction pipelines available
utilises linked data to answer a substantive question
SODA
DELTA
utilises web browser data to answer a substantive question
Conor's Fellowship
+ demonstrates proof of value on a case
ESRC SDA
Marta's grant proposal

Thank you for your attention



Follow ups

- Original purpose of DIGISURVOR was to test the feasibility of converting 3 existing datasets that combined survey and DTD to shareable and re-usable resources. We have/will produced list of variables, code and user guides to enable other researchers to develop these resources. Platform specific to X and URL domains. No guarantees of anonymisation. Need for researchers to do their own risk assessments.
- Demand for basic user guide in the methods, tools and best practice used in the collection, analysis, archiving and sharing of these datasets.
- Keep populating DIGISURVOR Github repository – share documents /code/outputs,
- Promote our email list to share updates, news, conferences, jobs, grant calls relevant to linked survey and DTD
- Building a network of interested researchers from our group. Look for funding to develop this initiative. Possible targets - EPSRC Network grant.
- Keep track and coordinate input into consultations /task force at UK and EU level that are pushing for researchers to get access to social media / online information service providers.
- Advertise training initiatives. DIGISURVOR working with the UKDS training and skills team to develop series of webinars that will introduce research community to working with linked data, and specific applications to social media and web browser data.

Infrastructure supporting collection, analysis, archiving and re-use of linked Survey DTD

- Decentralised /distributed projects collecting linked DTD and survey data. E.g DiCED, LSOM
- Data donation resource centres – software provision / open source to enable and support data linkage. e.g Next Platform
- 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, GESIS panel.dpd
- 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