DIGISURVOR Workshop: Summary & Next steps

Different models of DTD and survey data linkage collection and access

There are a range of models for DTD and survey data linkage collection, access and archiving. At the smallest scale are individual projects, these then scale up to nationally provided infrastructure. The 5 models with some examples of each variant are outlined below.

(1) Decentralised /distributed projects collecting linked DTD and survey data. E.g DiCED
(2) Data donation resource centres – software provision / open source to enable and support data linkage. E.g D31
(3) Platform provided option SOMAR at ICSPR/Michigan – again vetting process and 'clean room' built to allow for approved researchers to use the data.
(4) Commercially provided option panels – proprietary /paid for access. Managed in house, ethical compliance systems internal e.g. IPSOS
(5) 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

Open research data challenge – where does Digisurvor fit?

All of the above models face challenges in terms of converting the augmented data to allow for secondary analysis by external users, and do not have a plan so far as we know as to how these data can be made fully open access. The main constraints can be seen as at least four-fold. Firstly, maintaining respondent privacy and anonymity, and gaining informed consent are key to the success of such a project. However, creating new data that is sufficiently interesting and nuanced or useful to enhance the researchers purposes is also critical. There are also the issues of underlying bias and the integrity and quality of the new DTD based variables to consider, they cannot simply be linked and released. Some diagnostics and corrective measures are likely to be required. Finally to produce the data well as the technical complexity and skills required for the production of the variables.

How far are the current models set up to address the challenge of sharing / open release of linked survey and DTD?

	Open Research	2ndary analysis
Model 1	x	x
Model 2	x	x
Model 3	x	√
Model 4	x	x
Model 5	x	√

Model (1) the distributed model is typically not focused /resourced to enable open dataset production. It is only the data owners that have access to the datasets.
Model (2) is not designed for data collection or archiving/repository and thus doesn't directly confront the task of conversion of DD into shareable format.

Model (4) while it generates linked data is not under a funding obligation to provide those data for open access and arguably commercial incentives mean that it is unlikely to prioritise this.
Models (3) and (5) the platform partnerships and independent national non-profit Centres can develop secure rooms that allow for a limited pool of approved researchers to access the data. This works to open up the data for an extent but still operate limited system of access, privileged access for approved researchers. This type of approach also does not address the growing requirements of Journals for datasets to be deposited for re-analysis. Also datasets collected 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. An example of this for the 2020-22 ANES Social Media study under Model (3) can be found here https://electionstudies.org/data-center/2020-2022-social-media-study/ and the codebook of variables produced https://electionstudies.org/wp-content/uploads/2023/06/anes_specialstudy_2020_socialmedia_restricted_use_facebook_data_codebook_20230602.pdf

To date, this appears to be the only 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.

However, it is quite limited in the variables that are extracted so while they meet the anonymity criteria, whether they meet the utility or usability in full is questionable? Also there is no obvious attempt in this public release dataset of criteria being used to measure the robustness of that data – bias detection and correction. No code is released to allow others to replicate the work.

Summary

This workshop helpful to locating DIGISURVOR project in that space.

- Currently data sharing and open access are an end point in the cycle and largely an after thought rather than a primary concern.
- Our project is likely to recommend this thinking is built into or baked into the linkage process. i.e all models 1-5 need to consider what types of standardized and anonymized data should data producers generate as part of their project so that they can share it?
- If we think about current users for these linked data it follows a tiered or hierararchical model of access.
  - Inner core data owners/controllers;
  - Approved users in controlled environment
  - Approved users in uncontrolled environment; DIGISURVOR focus
  - Fully open data

Next Steps for the Group/Building the network

Github repository – share documents /code/outputs e.g. MIV variable list coded structural/substantive/validating, and by complexity, reliability and usability; conference paper;
Workshop No.2 12 months time – report back on our progress and your progress!
Building a network of interested researchers

Key takeaways from group discussions of 4 questions

1. Do our variables capture the main characteristics of interest regarding individuals’ Twitter use?

- Are respondents in Echo chambers or filter bubbles?
- More on topics (within politics, and outside-politics)
- Personal vs. work focus of activity
- Moderation activities, are they admins or moderators, do they report content, contribute to community notes,
- Visual content of posts – memes
- How topics are framed they are exposed to.
- Use of gendered communication style

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

- Caution: What is “political” could be very sensitive to definition of “political”
- Could consider Reliability, Complexity, Usability as additional coding

3. *What other variables might be included that we have currently not identified?*

Accounts followed:

- specific national, sub-national politicians; number of politicians followed
- Specific well known ‘influencers’
- Regional variation of accounts followed? in country and out of country
- Identify key accounts (e.g., Fox) for following, providing information on users who don’t post; Code them by topic
- This paper has a list of the top accounts followed by US Twitter users, with a categorization in type of account (politics, entertainment, etc.), also a useful list here from this paper
- Maybe use this method to get a measure of SES from the accounts followed

Include other fields:

- Sociology
- Economics
- Find key users from other fields

Wide range of non-political substantive areas of interest

- class,
- identity group membership,
- network heterogeneity
- geographical proximity, cosmopolitan/local
- health
- personality traits
- topics they post on /are interested in

Measures of political interest

- from follower patterns (members of congress, news sources)
- Out-party mentions,
- group polarization,
- attitudes to size of government, taxation

Use of LLMs/Gen AI

- Claud, Chatgpt, Copilot
- AI Literacy measure

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?

More general take aways:

- SNA methods could used to create new variables measuring respondent proximity /engagement in echo chamber/filter bubbles

- Future proofing – how time resistant are our measures

- How far do they and should they have cross-national application? Countries will have differing rules on the definition of PII and how it can be used. E.g. US vs Europe. Political views are sensitive data under GDPR. Should our measures aim for the most standardized to avoid problems. Or be open to others to decide based on their own national regulatory requirements.

- Misc –Other sources to cross check exposure to dis and misinformation - SNOPES

- How to ensure consent from respondents that data provided in original linked data to be reused in open data sets of for secondary analysis