UK DATA ARCHIVE 17/4/2018

- 8000 social science survey datasets (inc census).
  New domains like IoT energy hold the tantalising prospect of cross-disciplinary linkage and "collective intelligence"

- Broadly speaking, synthetic data is not an approach researchers want to adopt in the social sciences domain. Historically, we mitigate sensitivity of data by establishing a direct correlation between (perceived) disclosure risk and physical access controls.

- An accelerated move from a file-centric approach to IoT streaming data and very large graphs.

- Manual ad hoc application of de-identification is not sustainable in the long term, especially as we scale up to millions of IoT devices, nor is manual output checking.

- Government and administrative data currently locked up because of nervousness about disclosure. This is a loss to public policy making. When linkages are permitted, it can take literally years for a 3rd party data broker to produce a research dataset.

FUTURE

We need to develop:

1. Formalised descriptions of de-identification operations and attributes of sensitivity (e.g. age is usually less sensitive than sexual preference) which will ultimately support
   (a) RISK PROFILE OF LINKAGE OPERATIONS INTERNALLY AND AS REMOTE/FEDERATED LINKAGES. We cannot begin to operationalise linkage at scale and create appropriate algorithms without robust semantics about disclosure risk in the source data.
   (b) MORE ROBUST PROVENANCE CHAINS
   (c) IMPACT ASSESSMENTS ON UTILITY e.g. top-coding
   (d) MORE EMPIRICAL ACCESS CLASSIFICATION: there is a direct relationship between risk profile and Access Category

2. A taxonomy of "linkage" operations. Joins, Unions, Pivots, by geo, by concept etc etc.
   In conjunction with the above, this should make disclosure risk easier to "calculate".

3. NB - We have a tendency to overlook that disclosure is not just about individuals, it can be organisations, commercial entities or groups of individuals

4. More formal modelling of consent withdrawal. GDPR has some serious and nasty consequences for (a) citation and reproducibility of existing analyses (b) integrity of preservation artefacts.