Methods of Disclosure Control: The UKDS approach to review

Louise Corti
Director Collections Development
UK Data Service

Methods of Disclosure Control
CLOSER Knowledge Exchange event,
London, 18 January 2017

UK Data Service

Acknowledgement

To our ingest team for contributions to this presentation

- Sharon Bolton and Kay Eastaugh – survey data curation gurus

Our approach to Input Disclosure Review

- Advice: help data depositors make decisions about the relative risk in microdata they wish to share, and document these

- Work: undertake in-house disclosure review checks

- Outcome: data can be made available under various conditions, so sharing channels can be tailored to relative disclosure risk

Our responsibility

Help meet ethical and legal obligations

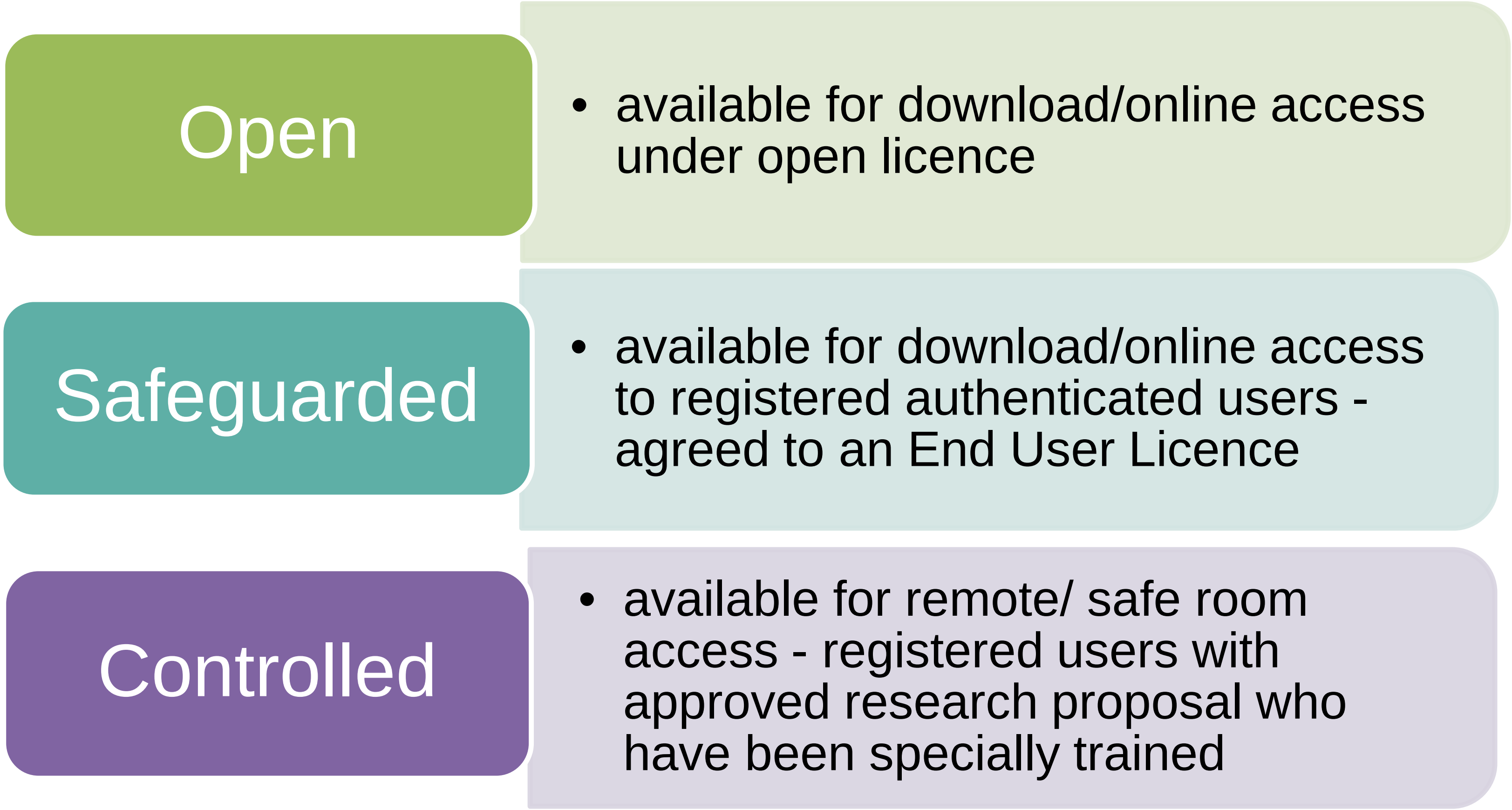
- Obtain informed consent for data sharing and long-term preservation
- Protect identities when promised
- Regulate access where needed (all or part of data) e.g. by group, use, time period

Protecting confidentiality: the ‘5 Safes’

- Safe data - treat the data to protect respondent confidentiality
- Safe people - educate researchers to use data safely
- Safe projects - research projects for ‘public good’
- Safe settings - SecureLab system for sensitive data
- Safe outputs - SecureLab projects outputs screened

5 Safes Video

Access spectrum



Open

- available for download/online access under open licence

Safeguarded

- available for download/online access to registered authenticated users - agreed to an End User Licence

Controlled

- available for remote/ safe room access - registered users with approved research proposal who have been specially trained

What data goes into which category?

- Most producers use own techniques for assessing ‘risk’ of identification - based on their acceptable thresholds

- Some use formal Statistical Disclosure Control (SDC) techniques to reduce the risk of disclosure to an ‘acceptable level’

- Most we speak to SDC takes an ‘intruder’ view, so that it is presumed that the intruder does cannot recognise anyone of the sample (e.g. family)

Not an exact science

- No magic formula to help us judge ‘objective’ risk
- We cannot give a one-size-fits-all rule book
- Recommend existing best practice for surveys

Front line guidance: ONS

- *Disclosure Control Guidance for Microdata Produced from Social Surveys (Oct 2014) with case studies*
- United Nations Economic Commission for Europe: Managing statistical confidentiality and microdata access
- ICO data privacy guidance: Conducting Privacy Impact Assessments: Code of Practice

We follow ONS Guidance on SDC

Assess disclosure risk based on three groups of potentially disclosive or Classifying variables

- Direct identifiers
- Key variables

  variables that, in combination, can be linked to external information to re-identify respondents in the released dataset. “Implicit identifiers” or “ “quasi-identifiers”
- Non-identifying variables
- Sensitive variables

Direct identifiers

Not usually found (on purpose) in data we receive

- Names; addresses; telephone numbers; email addresses; photos; (perhaps) IP addresses

- Unless explicit consent obtained for sharing, remove direct identifiers from data
- Securely store personal or sensitive data (separately)
- Store longitudinal linkage keys separately (to link admin/personal data and anonymised files)

Indirect identifiers

- Sensitive information: health information/medical conditions; illegal behaviour, drug/alcohol use etc.
- ‘Less sensitive’ information: age/birth date, specialist employment, religious affiliation, large household size, unusual health condition, geographic area
- Local specific characteristics
  - Household or community level e.g. flushing toilets, glazed windows
- Other text/string variables – too detailed
- Linked information - demographics in combination (e.g. demographics + geographies)

What we expect

- Treatment process to be as well documented as possible
- Which variables have been treated and how

- Good information through data documentation reduces user queries! Documentation is king!

- Examples:
  - Opinions and Lifestyle final check spreadsheet – reduces errors
  - Documentation to show variables included in different versions
  - Short report on disclosure treatment

Good documentation

- OLS – sent with the data

Cycle	Module no.	Client	Archive type	vars deleted/ amended	Serial number anonymised	Rage and DVSize top coded	Cases removed
Jan, Feb, April 2015 merged dataset	MAZ	ONS	EUL	DVAge3 NumPass citizen AZ_25 topcoded for purchases over 5k	Yes	Yes	None

- Change in survey managers? Need procedures!

UK Data Service

Checking - common techniques we use

Qualitative

- Look at univariate frequencies – low values for ‘risky’ variables
- Cross tabulate ‘risky’ variables to find small cell counts
- Choose thresholds, e.g. may be no cell counts <10 (ONS) or 30 (others)

Treatment

- Common: variable(s) – banding, top coding, reducing precision, remove variable, microaggregation
- Less common: adding noise. record swapping, simulation

Examples 1: ONS Wealth and Assets Survey

ONS longitudinal survey - Great Britain

- Wave 1 (July 2006 – June 2008
- Follow-up wave 2 (July 2008 – June 2010)

Looks at change in household assets change over the life course

- Data released in 2012 to UK Data Service for use under Special Licence
- Due to demand EUL also created

Risk Assessment

- Sample size
  - Wave 1: 30,000 household interviews
  - Wave 2: 20,000 household interviews
- Survey is a longitudinal, household survey
- Potential for extreme outliers on wealth variables

This information used when determining key variables:

where a combination might enable identification of an individual or household or an attribute relating to the individual or household

- Geography
- Country of Birth
- Ethnicity
- Religion
- Sexual identity
- Age
- Household Size
- Occupation

Applying Disclosure Control - EUL

- Remove households of size 10 and above
- Top code Individual Age at 80
- Give special consideration to the Wealth variable
  - all variables relating to wealth and finance top-coded
  - compromise - variables of lesser research importance removed to reduce the risk of identification
  - to retain full detail of the financial variables some rounding at the top level was still required

Additional disclosure control - EUL

As longitudinal dataset:

- Remove Geography from the EUL dataset
- Remove sensitive and 'observable' socio-demographic variables - country of birth, ethnic group, religion and sexual identity
- Recode combined age (HRP + spouse) Age into 5 year age bands
- Limit SOC (Occupation) to 2 digits
- Remove any flags that can identify births
- Suppress Wealth to three significant figures
- Top code Number of cars 4+

Reflection

- Removal of geography significantly reduces risk
- Data longitudinal but not pre-linked
  - Analysts need to link Waves themselves - extra step likely to reduce the likelihood of identifying split households and disclosing information about new household members
- Disclosure risk decreased due to age of the data
  - Wave 1 up to 6 years old; Wave 2 up to 4 years old.
  - Difficult to positively identify an individual from 10 year old data
- Data reviewed on a wave-by-wave basis to ensure the rules are still appropriate with 'evolving' data

Our recommendations

- Aggregate categories to reduce precision
- Top/bottom code or band ages continuous variables

  e.g., incomes, expenditure to disguise outliers
- Generalise meaning of detailed text, e.g. occupation
- Use standard coding frames – e.g. SOC2010
- Document changes made
- Talk to other data producers

Attempt to apply optimal SDC techniques that reduce disclosure risks with minimal information loss, and preserve data utility

Semi automating input SDC

- In-house use of ‘intruder’ algorithms to detect identifying ‘risk’ in data - individual cases that might stand out
- SDC Micro and ARX
- Computation and estimation of sample and population frequency counts to identify unique observations violating chosen thresh-holds
- Example principle: if frequencies of cases violating 2-anonymity exceed 5% of all observations the key variables used in combinations may present high risk of disclosure

Example 2: versions of ONS QLFS

- Joint review by UKDS & ONS of QLFS Special Licence data
- Assess potential for wider release of more detailed data at EUL
- How can SL data be treated to reduce risk to suit wider release without unacceptable loss of detail?
- Mitigate increased demand for Secure Access

- UKDS - data analytical risk assessment
  - Excludes external information
  - Examination of key variables and unique records against data intrusion simulation (DIS)
  - ARX software used
- ONS - penetration/intruder testing

Variables of interest - LFS

- Instances of several variables that cover the same concept at different levels of detail
- EUL - include only the least detailed categories rather than much banding/topcoding
- Birthdate
  - EUL - year of birth
  - Secure - month, day and year
- Industry code
  - Secure - 5-digit subclass for main, second, and last job
  - EUL - 4-digit industry class for main job only in EUL
  - EUL - 1, 2 and 3-digit for second and last jobs
- Geography
  - EUL - Region level
  - Secure - LA, NUTS3/4, Census Output Areas, Wards, parliamentary constituencies, Travel-To-Work-Areas etc.

Example 3: Health and Demographic Surveillance Systems (HDSS)

- Field sites observing the life events of 3 million+ people in 20 LMICs in Africa, Asia and Oceania
- Eg INDEPTH Wellcome Trust, NIH, and EC-funded
- UKDS collaborative work:
  - Ghana Millennium Villages study - DFID
  - Agincourt HDSS site, ZA .UKDS-DataFirst project (87,000+ people, 14,000+ households, 26 villages in semi-arid rural NE, since 1992)

HDSS Challenges

- Huge investments, multiple stakeholders
- HDSS face challenges in providing timely data
- Data sharing mandated…
- Often only summary demographics released
- But little other data available for social and economic researchers to exploit, without personal request
- Issue: disclosure risk and undocumented files
- Often no longer-term solutions for data access

- More capacity needed in data management and data preparation

Example 3a: Millennium Village Impact Evaluation in Northern Ghana, 2012

https://discover.ukdataservice.ac.uk/catalogue?sn=7734

- Millennium Villages Project (MVP) 'proof of concept' project to support African rural communities in meeting the Millennium Development Goals (MDGs)
- UK Department for International Development (DFID) provided a grant of £11.5 million to implement a new Millennium Village in northern Ghana
- GhanaMV - 2012 to 2017 with interventions targeting a cluster of communities with a total population of 26,000

Ghana MV data sharing

- Prospective data collection put at risk as no data shared
- PIs worked with UKDA to solve stalemate
- Disclosure risk assessment; post-hoc US IRB approval
- Difficult to gain trust in our data sharing procedures by data collectors/owners…
- Formats hard to review, process & analyse – 130 separate Stata files
- Little metadata in files; complex subfolder structures; poor documentation; little cross-referencing

Disclosure Review

- Identified potentially disclosive variables within each dataset as well as between groups of datasets
- Initial screening of data files for:
  - direct identifiers
  - key variables to identify individual units
- Frequency analyses of all variables across all data files to determine:
  - low-frequency responses and extreme outliers

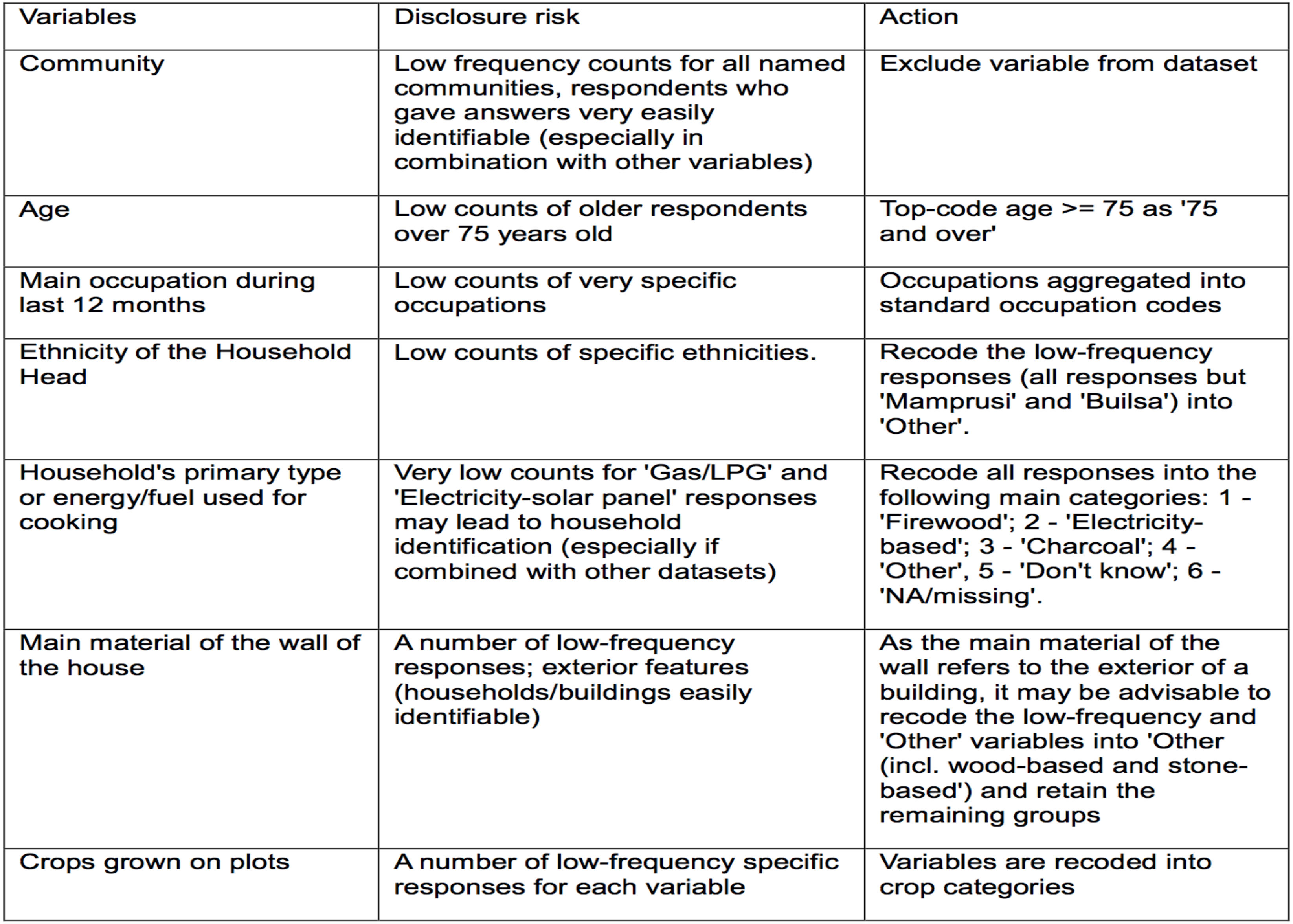
Assessment: semi-automated help

- Aim: ensure risk of linking confidential information with individual respondents was significantly lower whilst retaining utility
- *R sdcMicro* used to compute the sample and population frequency counts
- Frequency analysis tested whether responses to the combination of selected key variables were unique for any observation
- 162 observations identified where the combination of key variables was unique for those individuals

Variables assessed

- Granular and direct identifiers:
  - raw age, community and village names had very small frequency counts - excluded from dataset
- Those for which local knowledge is essential to indicate risk - implicit or quasi-identifiers
  - ethnicity, fuel type use, toilet facilities with flushing mechanisms, house wall material – recoded/grouped
- See UKDS-ESPA Guide: Sharing social data in multidisciplinary, multi-stakeholder research

Household survey variables assessed



Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables)	Exclude variable from dataset
Age	Low counts of older respondents over 75 years old	Top-code age >= 75 as '75 and over'
Main occupation during last 12 months	Low counts of very specific occupations	Occupations aggregated into standard occupation codes
Ethnicity of the Household Head	Low counts of specific ethnicities.	Recode the low-frequency responses (all responses but 'Mamprusi' and 'Builsa') into 'Other'.
Household's primary type or energy/fuel used for cooking	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	Recode all responses into the following main categories: 1 - 'Firewood'; 2 - 'Electricity-based'; 3 - 'Charcoal'; 4 - 'Other', 5 - 'Don't know'; 6 - 'NA/missing'.
Main material of the wall of the house	A number of low-frequency responses; exterior features (households/buildings easily identifiable)	As the main material of the wall refers to the exterior of a building, it may be advisable to recode the low-frequency and 'Other' variables into 'Other (incl. wood-based and stone-based') and retain the remaining groups
Crops grown on plots	A number of low-frequency specific responses for each variable	Variables are recoded into crop categories

UKDS access solution

- Release 1: Household data only
- Special Licence condition
- Proposed Data Access Committee and procedures for decision making about applicants
- And how access to more than one dataset is to be judged (e.g. household data plus bloods)
- For ease of access administration, each conditional Special Licence (bloods, anthropometrics), is held under a separate study number, especially if access to one of the data collections precludes access to another

New life for HDSS data: beyond demography

- Recent complete restructuring of unavailable Agincourt HDSS data to meet social science needs:
- Linked panel data format (long form) at 3 levels:

Individual level data (N=200,000)	**Life events from 1992 - every person** **Educational events from 2000 - most people** **Labour force events fom 2000 - most people**
Household level data link to Person ID	Size by year from 1992 Assets and consumption from 2000
Village level information	

- Secure access only
- Exemplary showcase for release of complex data

Summary: review and access control

- Balance between protecting respondents' confidentiality and maintaining research utility of data
- Open where possible, closed when necessary
- Combine anonymisation with access control to preserve usability - create multiple versions of data
- Accept that some research can only be done with identifying data e.g. research on patients with specific diseases
- Go back to the 5 Safes – consider sharing via an accredited Secure Lab or Secure Research Data Centre (ISO27001)
- Producers benefit from providing clear documentation on disclosure review and treatment!

Contact

Louise Corti and teams

UK Data Archive

University of Essex

Colchester

CO4 3SQ

corti@essex.ac.uk

Collections@ukdataservice.ac.uk