ENABLING ACCESS TO UK SOCIAL SCIENCE DATA ACROSS THE DISCLOSURE SPECTRUM

LOUISE
CORTI

ASSOCIATE DIRECTOR
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

IDCC
AMSTERDAM,, 15-16JANUARY 2013

UK • DATA ARCHIVE

WHO ARE WE?

- an “umbrella” data service organization
- provides infrastructure and shared data services for:
- UK Data Service now incorporating
  - Census Support and Secure Data Service

- Providing access to social science data created by government and academics for use by government, academics, commercial researchers, etc.

- Data are free at the point of use for most users

OPEN DATA

Definition of ‘open’

“A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike.”

Open Defitniton, http://opendefinition.org/

THE CASE FOR & AGAINST OPEN DATA

Many good reasons:

- Innovation, enterprise, efficiencies, accountability, verification
- EU strategies: Environmental Information Regulations (2004) and EU INSPIRE compliance (2007)

Negatives:

- cost and time for preparation
- low data quality, poor data formats, low value, misuse
- loss of data integrity
- may lead to closure/removal of more significant data

UK OPEN GOVERNMENT LICENCE

- For Information Providers in the public sector
- Free use and re-use for all purposes, both commercial and non-commercial
- Allowed: Copy, publish, distribute and transmit; Adaptation & Commercial Exploitation
- Caveat: Acknowledgement & No breach of DPA or Piracy/Electronic Communications Regulations

LAWS IN THE UK GOVERNING PERSONAL DATA

Data Protection Act, 1998

- Data that contains personal, confidential or sensitive data
- Information on business, income, health, medical details, and political opinion, offences

Statistics and Registration Service Act (SRSA), 2008

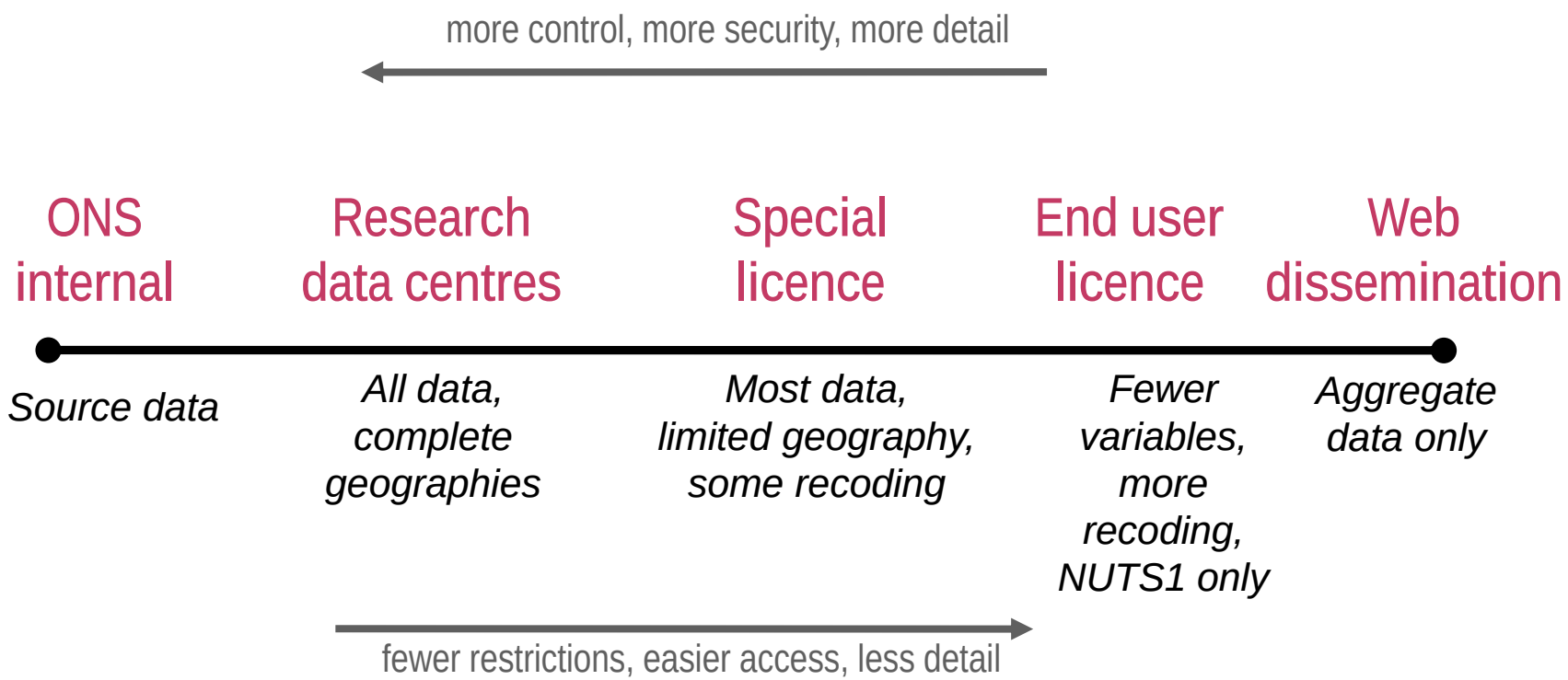
- allows information sharing between public authorities and the Statistics Authority for statistical purposes
- confidential information should not be disclosed by anyone; unlawful disclosure is a criminal offence punishable by a fine or imprisonment

The public benefit of data access should be greater than the harm or distress caused by disclosure

OUR LICENCES & CONDITIONS OF ACCESS

- Depositor's Licence - gives us right to copy, archive and distribute for agreed use
- DepositorAgree End User Licence (EUL) and registration
- Special Condition datasets, e.g. depositor permission required for use or publications; HE use only
- Special Licence datasets, e.g. ONS Approved Researcher access, British Crime Survey: low-level geographic data
- Secure Data Service datasets – data cannot be downloaded – access only via secure area, training required, additional username/password issued

OPEN DATA SPECTRUM



From Ritchie, 2006

ALLAYING FEARS & ENSURING TRUST

- Our two main concerns:
  - Fear of disclosure leads to loss of detail in published data
  - Fear of disclosure leads to ‘classification creep’
- Have to ‘prove’ safety (i.e. its not us who leaves data sticks lying around!). Trusted Digital Repository and ISO27001
- Able to provide access to disclosive data via this trusted relationship, producer vetting and penalties for breaches
  - low-level geographies, single age groups, detailed industry and occupation codes

Safe use = Safe projects Safe people Safe data Safe setting Safe outputs

ANONYMISING QUALITATIVE DATA

- No qualitative data with secure access yet
- Some studies still require depositor permission
- Anonymisation undertaken by the depositor
- Follow our Guidelines and advice
- But still need to read all data as risks too high for our reputation..very time consuming…expensive
- Will get back to depositor with advice

KEY POINTS FOR ANONYMISING

- never disclose personal data - unless consent for disclosure
- reasonable/appropriate level of anonymity
- maintain maximum meaningful info
- where possible replace rather than remove
- identifying info may provide context, do not over-anonymise
- re-users of data have the same legal and ethical obligation to NOT disclose confidential info as primary users

SOME BASIC RULES

- remove direct identifiers

  names, address, institution, photo
- reduce the precision/detail of a an entity through aggregation

  birth year vs. date of birth, occupational description, area rather than village
- generalise meaning of a detailed entity

  educational background/level
- restrict upper lower ranges of a variable to hide outliers

  income, age (over 75)
- combining data

  creating non-disclosive rural/urban classification from place names

OUR STRATEGIES FOR ANONYMSIATION

- plan or apply editing at time of transcription

  except: longitudinal studies - anonymise when data collection complete (linkages)
- avoid blanking out; use pseudonyms or replacements
- avoid over-anonymising - can distort data, make them unusable, unreliable or misleading
- consistency within research team and throughout project
- identify replacements, e.g. with [brackets]
- keep anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files
- xml mark-up can be used for anonymisation

  <seg type="anonymised">word to be anonymised</seg>

ANONYMISING QUALITATIVE DATA

Example: Anonymisation log interview transcripts

Interview / Page	Original	Changed to
Int1		
p1	Spain	European country
p1	E-print Ltd	Printing company
p2	20th June	June
p2	Amy	Moira
Int2		
p1	Francis	my friend

AUDIO-VISUAL DATA

- Digital manipulation of audio and image files can remove personal identifiers

  e.g. voice alteration, image blurring (e.g. of faces)

- Labour intensive, expensive, may damage research potential of data

Better:

- Obtain consent to use and share data unaltered for research purposes
- Avoid mentioning disclosing information during audio recordings

WHO SHOULD DO THIS WORK?

- Ideally researchers themselves
- Various tools to help, ICPSR and IQDA
- For deposit in institutional responses, for human data, checks must be carried out by independent person
- Even if participant gives permission for all their contribution to be used as is, advise checks!
- Repository may be at risk from libel/slander accusations
- For ‘secure’ data, disclosure control checks on all outputs

NEGOTIATED ACCESS CONTROLS FOR TIMESCAPES

Definitions for Levels of Access to Timescapes data				
Type of use/user	**Key purpose**	**Examples of data available***	**Authentication system**	**Requirements for use**
Public	Showcase data on public areas of LUDOS & Ts websites	Metadata & anonymised "taster" research data	none	Email and contact details
Registered users	Data sharing and reuse for registered users	Anonymised project data; some unanonymised data with participant consent, e.g., images, video; researcher notes	Database of user accounts	Authentication; user registration; and sign end user licence**
Approved users (Case-by-case)	Registered users access sensitive data subject to vetting by Ts team members or designated representative	Disclosive data, unanonymised data, visual and audio data	case-by-case review of individual applications; plus database of user accounts	User application reviewed; and authentication; User registration and sign end user licence
Embargoed data	to enable preservation of data too sensitive for sharing now, and to enable data to be shared at later dates.	most sensitive data; data with ambiguous consent AND with researcher approval	not applicable	not applicable

FUTURE RESEARCH DATA ACCESS STRATEGY

- Clear guidance on classification of data
  - risk of and impact of disclosure
  - risk of and impact of organizational reputation
- Use anonymisation techniques appropriately – have multiple “versions”
- Introduce appropriate access methods for non-Open Data
- ‘Graduated’ licences between producer and supplier, and supplier and user
- Cost-effective authentication/authorization

CONTACT

UK DATA ARCHIVE
UNIVERSIY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX CO4 3SQ

T +44 (0)1206 872145
E corti@essex.ac.uk
W www.data-archive.ac.uk