Documenting your data

Dr Sharon Bolton
Data Curation Manager

BD013 – Managing, Curating and Publishing Data
IADS Big Data and Analytics Summer School
University of Essex, 25 August 2015

UK Data Service

Documenting your data

- Why document?
  - Research sustainability and long-term data usability
- Study-level documentation
  - Research context and design, data collection methods, data preparation and results or findings
- Data-level documentation
  - Labelling and documenting individual data items - numerical, textual or audio-visual - embedded in data or recorded elsewhere
- Catalogue metadata
  - Describes purpose, origin, time, geographic location, creator, terms of use - UK Data Service deposit form. Helps user decide whether this is the right dataset for their research

Good documentation

- context of data collection: history, aims, objectives
- data collection methods: protocols, sampling design, instruments, data scale and resolution, temporal and geographic coverage
- structure of data files, cases, records, variables and relationships between files
- data sources and provenance of materials - transcribed or derived data
- data validation, checking, cleaning, other QA procedures
- modifications made over time since original data creation
- changes over time to methodology, content, labelling, measurements , sampling
- information on confidentiality, access and use conditions

Documentation & metadata formats

- Often deposited as PDF – technical reports, questionnaires, findings reports
  - Popular, free readers, well-supported
  - Not ideal for archiving, too software-dependent
- Word and RTF – same as above.
- XML, with schema - Data Documentation Initiative (DDI) metadata can be shared across systems easily.
  - CLOSER project questionnaires can be read direct into Nesstar

Documentation rights

- Think about structure and contents of your documentation
- Does an academic paper or a published monograph form part of your documentation?
  - Publisher may have copyright – no redistribution
  - Be careful how much you quote
- Make sure the essential information on how to use and analyse the data is available outside copyrighted articles

Documenting big data?

- Data may not have been produced for sharing – may be limited information available
- Data formats/access method might be different, but users still need information
- Copyright may still need consideration
- Weblink/API - what kind of 'catalogue record' information will users need to see?
- How user might reasonably subset data – what would they need to know?