

---

# Data Formats

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

---



---

# Formats: Social Science Data

- If you want others to use your data and documentation, think about the formats you want to produce
- Popular, well-supported software packages
- Platform independent
- Long-term preservation?



---

# Formats: Survey microdata

- UK Data Service:
  - SPSS
  - Stata
  - Tab-delimited or CSV formats
- Can be read into R – use the 'foreign' library
- Tab-delimited and CSV – can be read into databases, Excel (within limits)
- Metadata – what format? RTF, XML, text



---

# Formats: Macrodata

- UK Data Service:
  - UKDS.stat - OECD data warehousing technology
- Explore and analyse online
- subset and download data (CSV) to read into databases or Excel (within limits)
- Metadata (data-level documentation) extensive online, minimal in downloaded subsets



---

# Formats: Big Data

- What will users want? Online analysis and subsetting in Cloud or download facilities?
- CSV, SQL – R/R Studio, Hadoop, MongoDB etc.?
- Downloading smaller subsets – other analysis formats?
- Don't forget the metadata and documentation



---

# Formats: Documentation and metadata

- 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

