Cite as: Veerle Van den Eynden (2018) Exercise: Data publication. UK Data Service, UK

UK Data Service

Exercise: Data publication

Evaluate various options for publishing data and consider the pros and cons of each option:

- discipline-specific data repositories
- generic data repositories
- institutional repositories
- data journals

What would be the best repository to publish and disseminate the following datasets:

1. HIV Uganda survey data (SPSS, 423 responses, 250 variables) and collection of 38 interview transcripts (RTF format). Public access (survey); restricted access (interviews).
2. Airline safety performance data, replication data for paper “Profitability and Product Quality: Economic Determinants of Airline Safety Performance”, Journal of Political Economy. Data from 35 airlines for period 1957-2010, with accident/incidents data, operation data, financial data. Public access.
3. Midlife in the United States (MIDUS), a national longitudinal study (3 waves to date) of health and well-being from a national sample of 7000+ American adults aged 24-74. Date are captured by different protocols (comprising around 20,000 variables): survey measures, cognitive assessments, daily stress diaries, clinical, biomarker and neuroscience data. Restricted access.
4. Digitised (scanned) Derek Freeman anthropology field notes with annotations, Borneo, 1950s. Public access.

Consider for example the following repository / publishing options:

- ICPSR: http://www.icpsr.umich.edu
- UK Data Service ReShare: http://reshare.ukdataservice.ac.uk
- Figshare: http://figshare.com/
- Zenodo: http://zenodo.org
- Scientific Data: http://www.nature.com/sdata/
- Journal of Open Health data: http://openhealthdata.metajnl.com

First explore each of these repository / publishing options and consider:

- Can anyone publish data via these systems?
- How are datasets exposed / disseminated, so others can find them?
- Do dataset metadata give enough information to understand the data content?
- Can anyone download data or are there access controls in place? If so, why?
- What is the repository’s data use agreement (if any)?

Then decide which place would be best for each dataset above, and which criteria are important for your decision.