Collecting and using metrics: the UK Data Archive point of view

Matthew Woollard

Head of Digital Preservation and Systems, UKDA

Value and Benefits of Data Sharing and Management
Research Data Management Forum,
Manchester, 1 May 2009

UKDA

Brief

- How do we measure the value of a dataset?
- How do we measure the quality of a dataset?
- How do we measure the impact of a dataset?
- What metrics should the DCC be promoting / collecting to demonstrate economic and social benefits from effective data curation?
- How can we demonstrate to funders and policy makers quantitative evidence of the benefits of investment in data infrastructure?

UKDA

How do we measure value of a dataset?

- Value of dataset = value to original research
  - costs of creation
- Value of archived dataset = value to secondary research
  - costs of preparation
  - costs of retention
- Policy matters

UKDA

How do we measure quality of a dataset?

- Content
  - methodologies employed / rigour
  - data quality
  - originality
  - significance
- Ancillary material
  - Documentation / context
- Technical
  - file formats /software?
  - Submission details

UKDA

How do we measure impact of a dataset?

- Accesses
  - problems

    audience

    dissemination method

    aggregate figures hide heterogeneity of access

    aggregate figures hide non-use

    individual figures hide selection procedures
  - types of measurement

    period (days after release?)

UKDA

Number of days between publication of a study and its first use, 2003-2008

	median	mean	studies	studies not used
2003	36	182	112	3
2004	34	124	166	10
2005	15	67	139	2
2006	16	73	226	7
2007	17	65	189	5
2008	15	35	291	8

Note: caveats too numerous to mention in this space! See paper.

How do we measure impact of a dataset?

- Accesses <> uses
- Accesses poor measure of impact

- UKDA Data Accesses <> UKDA Accesses

  - SN 3955
    991 data accesses 1999-2009
    10,500 documentation accesses Jan-Jul 2007

- Dataset impact vs. service impact

UKDA

How do we measure impact of a dataset?

- Statistics
  - webometrics [links/references to datasets]
  - bibliometrics [links/references to datasets]
  - web analytics [usage of site]

- Evidence of use / service provision
- Benefit in terms of Quality Management and Quality Improvement

UKDA

How *should* we measure impact of a dataset?

- Potential audience / market penetration
- Use to which data is being put (research/training?)
- Publications relating to key datasets
- Enquiries about data (not about access)
- Attendances at (themed) data workshops

- Costs
- Accesses

What metrics demonstrate economic and social benefits from effective data curation?

1. Costs

- No clear benchmarks across domains
- Metrics also useful to users and providers

- Costs of delivery per access
- Cost of retention (per access?)
- Cost of replication/replacement

- Opportunity costs
- Productivity benefits
- Capture cost estimates at ingest

2. Quality of data

- Bibliometrics/webometrics
  - interpretation, time lag
  - research instruments not just data

- Value of data (counter-factual)

- Peer review (scholarly/technical)
- Specialised publication for data?

2. Quality of service

- Value of service
  - value-added
  - user support
  - training / teaching
  - promotion
  - migration / transformation
  - infrastructure
  - “trusted” status
  - staff knowledge

Last points

There are costs and impacts of the collection of impact data / metrics on any service

Be highly selective and sensitive to costs of any impact assessment