PERSISTENT IDENTIFIERS AT THE UK DATA ARCHIVE

MATTHEW WOOLLARD
DIRECTOR
UK DATA ARCHIVE
UNIVERSITY OF ESSEX
THEORY INTO PRACTICE

• Meanings of words
  • Persistence must mean enduring
  • Identifiers must be unique
  • Digital Object should be clearly defined to ensure appropriate granularity.
CURRENT SITUATION

- Test British Library/DataCite’s DOI allocator
- Building web services to align with above
- Agree strategy for phased implementation
WHY NOT READY YET?

• Archive “data collections” are not digital objects
• Desire to resolve inconsistent use of version/edition
  • Ensure they’re machine-actionable and not just human-mediated
• Integrate processes with digital preservation activities
  • Incremental changes are machine-controlled
• Current infrastructure / work flows
• Desire to “get it right” first time
VERSIONS AND EDITIONS

- Ugly terminology (compare OAIS AIP Edition/Version) with library science definition.
- “Archive” / “library” perception that citation must be perfect.
- Social science users want most recent (95%) for research and older for verification (5%).
- Approx 15% of all data collections altered within first year after first publication.
USER-CENTRED IMPACT

- Desire to move from “archive-practice” centred approach to user-centered approach.
- Impact is defined in terms of effect on users. It covers (amongst other things):
  - Changes to variables/data
  - Changes to documentation
  - Regrossing of a data series
  - ‘Waves’ in series
EXAMPLES

- Low impact
  - Correction of spelling in variable labels
  - Small changes in variable labels
  - Removal of (erroneously supplied) admin variables
  - Correction of spelling in metadata
  - Minor changes in documentation
  - New index terms
  - Additional documentation added (non-fundamental)
  - Change in access conditions
EXAMPLES II

• High impact
  • New variable added
  • New labels/value codes added
  • Weighting variables reconstructed
  • Wrong data supplied (e.g., March not April)
  • Mis-coded data (e.g., Don’t know/Refused confused)
  • Change in format (file migration)
  • Change in access conditions
RECONCEPTUALISATION

• Ingest to be understood as an ongoing dynamic process
• Concept of an instance to denote changed collection (and amenable to machine actions such as comparing checksums)
  • Internal change during ingest process (unreleased)
    => new internal instance
  • Low impact change (released)
    => new external instance with unchanged DOI
  • High impact change (released)
    => new external instance and new DOI
SOLUTION?

• Original solution
  • Phase 1: DOI allocated to core metadata (title, etc.) relating to a data collection
  • Problem: even titles can change

• New solution
  • Phase 1: DOI allocated to metadata relating to each external instance of a data collection
    • DOIs resolve to “jump” page pointing to all external instances (and indication of internal instances?)
    • New DOI = High Impact change, with explicit logging but we could also update an existing DOI with low-impact change information
  • Phase 2: All instance-specific data made available over time (where allowed)
Without whom...

Oscar Dovao
Kay Easthaugh
Hervé L’Hours
John Shepherdson