closer
The home of longitudinal research

Cross study and institution perspective

Jon Johnson

13 September 2019

CLOSER, UCL Institute of Education

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UCL

MRC Medical Research Council

Overview

- CLOSER Production processes
  - Archivist / Questionnaire capture
  - Mapping of topics
  - Dataset management
  - Improvement strategies

- Engagement with funders, strategic priorities

Question capture - Ongoing challenges

- Lack of reuse of questions across different questionnaires in Archivist slowed entry.
- Use of DDI Instance as the only import and export mechanism reduced flexibility.
- Large CAI without existing code
- Export should only include elements that are in the questionnaire, make sure there is not any extraneous content
- You need a way to *validate output*, content as well as a well formed document
- *DDI Fragments* may be a better strategy, less fragile and possibly easier to handle
- Multiple and new formats (when new studies added) to parse means continuous development

Import of questionnaire content reflections

- High quality web-based editor Archivist allowed:
  - Recruitment of non-technical staff
  - Reduced training times
  - Efficient metadata entry
  - Monitoring of progress and assistance to remote entry
  - Ability to copy whole questionnaires to edit
- Use existing code (e.g. Blaise) to parse then edit speeds up entry
- Imports are not perfect, so *editing is always* needed (QA)
- Online documentation improved consistency and training length
- Verification process improved QA
- Time taken to verify input should not be underestimated.

Mapping - variables and questions

- Many studies had maps in e.g. Excel between questions and variables
  - However, the naming didn't always make reuse of this easy
- For CAI questionnaires, this was a major time saver to use Excel
  - Allowed .txt import into Archivist for validation/manage conflicts
- Derived variables were treated as a variable to variable map, so even if we did not have the algorithm we at least knew what went into each derived variable.

Mapping - concepts / topics

- In a consortium of studies, agreeing on a set of concepts requires negotiation
- Finalised list was based on a merging of HASSET and MESH
- Two levels, broader and narrower
- Most studies did not have maps of variables to concepts
  - Doing this is time consuming
  - Inherited concepts from questions -> variables which sped things up

Dataset capture

- We used Sledgehammer as this supported DDI-Lifecycle 3.2
- Python scripts were used to tidy up and add some additional content labelling and reference to DOIs, location of dataset
- A minimum standard was set for output
  - Variable name, label, min, max, mean, code list
  - Optional frequencies, standard deviation
- The list of variables on datasets processed using Sledgehammer was loaded into Archivist to facilitate mapping

Dataset capture reflections

- It allowed studies to process the datasets without having to pass the data to CLOSER (we only handle the metadata)
- The software was simple to use by studies and gave a consistent output after guidelines and standard configuration files were used.
- Reduced in-house development

Improving the Processes

- Import tables of information for later editing
  - Question and code lists, conditions etc
- Use of DDI fragments from existing CAI code
  - via Colectica ingest from CAI code
- Capture of content from large semi-structured PDF's
  - Development of Machine Learning / NLP strategies
  - Import from this would use previous import mechanisms
- Validation of coverage of metadata content
- Quality assurance of dataset summary statistics

Data management in birth cohort and longitudinal studies (2016) *

Issues

- *Researchers will continue to demand updated metadata from new data collections and this will need to be supplied to CLOSER or its successors*" and that
- "T*he studies themselves should be able to utilise the metadata collection to enhance and improve their own data management processes. At present, most studies are not in a position to do this primarily because their data infrastructures are outdated*" and
- *"Alongside this, the data landscape is continually evolving and the demands on data management at the studies are increasing, but the technology to manage this is in general not keeping up with current or future requirements*

Pascal Heus (2016) Data management in birth cohort and longitudinal studies (Scoping Study)

Data management in birth cohort and longitudinal studies (2016)

Recommendations

- Transition from legacy database technology
- Roadmap to enable content to be migrated to ‘big data’ technologies
- Improve file management technology and practices
- Metadata Management
- Imaging Management
- Strengthen statistical disclosure control capacity

Data Landscape Review Recommendations*

1. The promotion of further and close coordination, both within individual councils and across councils, to enable a **common approach to commissioning, supporting longitudinal studies and the way that their data are accessed**.
2. The implementation of a common governance framework to ensure more **consistency between the processes used by different studies to access data**, or required by different data controllers.
3. The continued support for **trusted, certified and streamlined data access processes**.
4. Funders to ensure planning and support for **future-proofed and interoperable, shared dissemination and storage infrastructure and facilities for data assets**.
5. A better appreciation by funders of the **importance of consistent and interoperable metadata standards for longitudinal investments**.

Park, Johnson, Corti (2018) Longitudinal Data Landscape Review (internal report to ESRC)

Data Landscape Review Recommendations

6. The requirement for **funder data management and sharing plans to include more detailed information on metadata standards** to promote inter-operability and discoverability for future studies.
7. Improving **discovery, visualization and analysis offers for different users**, including academics and policy makers.
8. The initiation of a policy and strategy **for enabling documentation** of, and access to, important UK legacy data collections **once a large study closes** or the Principle Investigator (PI) leaves.
9. Ensuring that **impact from data assets can be better tracked and cited using persistent identifiers**, both for data and for users.
10. Funding a longitudinal data resource centre that can foster a **collective approach to the core activities involved in longitudinal data creation, access, use and impact tracking**.