From 1911 to 2013: Renewing UK Birth Cohort Studies Metadata

Jon Johnson
Centre for Longitudinal Studies, Institute of Education, University of London

Jack Kneeshaw
UK Data Archive, University of Essex
What is CLOSER? What is the CLOSER Uniform Search Platform (USP)

- Cohorts and Longitudinal Studies Enhancement Resources (CLOSER) is a consortium of nine of the UK’s leading birth cohort and longitudinal studies; it aims to maximise the use, value and impact of these studies: http://www.closer.ac.uk

- Cohorts are a mix of national and regional; long-established and in infancy; first participants in Hertfordshire Cohort Study born in 1911

- The Uniform Search Platform (USP) will provide a search interface, which will return relevant results for both basic and advanced metadata queries across the studies; it will also standardise practice across the participating studies

- **Caveat**: Other presentations are about tools and resources that are much more mature
Vision for the USP

- Portal to discovery of hundreds of thousands of variables, questions and data collection instruments across the nine longitudinal studies:
  - covering survey and biomedical data collection
  - promoting CLOSER harmonisation work
  - state-of-the-art searching tool
  - focus on improving visibility of associations between (currently) disparate metadata items
  - shared subject/topic classification

- We should remember that this is massively ambitious; something that matches or surpasses the best multi-study metadata repository out there:
  - RAND Survey Meta Data Repository covering the HRS family of studies: [https://mmicdata.rand.org/megametadata/](https://mmicdata.rand.org/megametadata/)
Why do it?

- Benefits to **users**:  
  - single resource discovery portal – replacing a fractured resource discovery landscape  
  - lowers barriers to conducting cross-cohort analysis  
  - increased visibility of cohort data and resources

- Benefits to **data managers**:  
  - standardised metadata management workflows – currently curated in isolation  
  - workflows in place for future ‘joiners’

- Benefits to **Principal Investigators/survey commissioners**:  
  - make prospective harmonisation easier  
  - promotion and re-use of tested questions and instruments
Assumptions, constraints

- Not a data repository

- Not a major software development project:
  - major £££ is for metadata creation/enhancement

- DDI-L agreed as standard for metadata exchange:
  - covers subject areas (bio and soc science) and data collection methods ('hard' instrument and survey)
  - designed for marking-up longitudinal/repeated metadata items

- Colectica Designer selected as preferred metadata ingest/editing software
Challenges

• Legacy metadata:
  • elderly and decrepit!
  • not always designed for equivalence within a study, much less across studies
  • differing or non-existent naming conventions
  • substantial (manual) effort required to establish equivalences and level of equivalence

• Metadata managed by five or six different units: different formats, workflows, vocabularies

• Relative lack of familiarity with DDI-L:
  • uneven knowledge across study units
metadata: state of play

• >200k variables

• c.150 data collections:
  • CAI, PAPI, nurse visit, clinic-based protocol, biosamples, etc.

• c.85 validated survey instruments
  • GHQ, AUDIT, Malaise Inventory, etc.
  • c.10 instruments used in >1 study

• c.20 validated clinical measures
  • blood pressure, bone density, lung function, etc.
  • range of instruments used

• c.15 cognitive or physical tests
How to do it?

- USP will be a web interface that sits on top of a central repository fed by metadata created and delivered both by the individual study units and the CLOSER core.

- Study units continue to curate metadata as they see fit; but not in conflict with proposed USP metadata profile.

- Substantial metadata creation and enhancement to be undertaken by the study units: inputting historical questionnaires; mapping between data items and data collection.

- CLOSER core responsible for identifying common (cross-study) variable and question schemes, allowing studies to reference these and also any agreed controlled vocabularies (concept, life stage etc.)
Study unit

- Curate/augment metadata compatible with agreed metadata model
- Clerical-level input of questionnaire metadata via Survey Editor
- Import / transform variable metadata via Colectica Designer
- Clerical-level input of clinical instrument / protocol metadata via Colectica Designer

Central repository

CLOSER core (+ WPs)
CLS DDI3 Survey Editor

- CLS-developed Ruby on Rails application to input questionnaire metadata; design goal was to create environment for clerical-level input of metadata from a questionnaire which can be output as DDI-L XML and later integrated with corresponding variable metadata

- Database design was modelled closely on the DDI XML schema and a Rails front end used to try and abstract as much as possible for the end user

- Input categories, create code schemes, add literal questions and create question components by combining these within if, loop and sequences constructs

- So far, inputted self-completion/face-to-face PAPI and simple CAI instruments for 4 studies, covering 12 data collections

- [http://www.youtube.com/watch?v=MNphgQDzfo4](http://www.youtube.com/watch?v=MNphgQDzfo4)
Colectica Designer

- Cleaning and reorganisation of metadata using Colectica Designer and SDK:
  - identifying and rationalising duplicate code and questions schemes
  - flagging orphaned data items, i.e. where data items have no corresponding code, question, instrument, derivation

- Each study unit and the core is treated as a separate Colectica instance within this federated system
Where are we now?

Stage 1: Project management and planning

Stage 2: Metadata enhancement tool(s) development (inc. pre-pilot)

Stage 3: Metadata enhancement tool(s) pilot

Stage 4: USP search interface specification and procurement

Stage 5: Metadata enhancement

Stage 7: USP search interface development and testing

Stage 8: USP search interface beta release

Stage 9: USP search interface communication and promotion

Stage 12: USP search interface further development

Stage 13: USP search interface release
What next?

Stage 1: Project management and planning

Stage 2: Metadata enhancement tool(s) development (inc. pre-pilot)

Stage 3: Metadata enhancement tool(s) pilot

Stage 4: USP search interface specification and procurement

Stage 5: Metadata enhancement

Stage 7: USP search interface development and testing

Stage 6: USP search interface beta release

Stage 9: USP search interface communication and promotion

Stage 12: USP search interface further development

Stage 13: USP search interface release
See you next year ;-) 

- http://www.closer.ac.uk - CLOSER
- http://www.youtube.com/watch?v=MNphgQDzfo4 - CLS DDI3 Survey Editor
- kneejw@essex.ac.uk - me

- I have leaflets and I don’t want to take them home!