closer
Cohort & Longitudinal Studies Enhancement Resources

CLOSER and the Challenge of Historical Metadata

Gemma Seabrook

MRC Medical Research Council

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

CLOSER

- 5 year collaborative project
- Maximise the use, value and impact of cohort and longitudinal studies
- Based at UCL Institute of Education

closer
Cohort & Longitudinal Studies Enhancement Resources

CLOSER Studies

- Studies:
  - Avon Longitudinal Study of Parents and Children (ALSPAC)
  - British Cohort Study (BCS70)
  - Hertfordshire Cohort Study (HCS)
  - Life Study
  - Millennium Cohort Study (MCS)
  - National Child Development Study (NCDS 1958)
  - National Study of Health and Development (NSHD 1946)
  - Southampton Women's Survey (SWS)
  - Understanding Society

closer
Cohort & Longitudinal Studies Enhancement Resources

CLOSER Partners

- Other partners:
  - UK Data Archive
  - British Library
- Funders:
  - Economic and Social Research Council
  - Medical Research Council
- Also working with Colectica and Metadata Technologies

closer
Cohort & Longitudinal Studies Enhancement Resources

Work Streams

- Data Harmonisation demonstration projects
- Data Linkage demonstration projects
- Research Impact
- Training and Capacity Building
- Uniform Search Platform

closer
Cohort & Longitudinal Studies Enhancement Resources

Current situation

- I want to know about smoking in pregnancy and effects on income in later life
- Want to look at ALSPAC, NCDS, NSHD, BCS and predictively at MCS
- NCDS, BCS, MCS – UKDA
- ALSPAC and NSHD – contact studies

closer
Cohort & Longitudinal Studies Enhancement Resources

The (near) future

- Search for smoking on CLOSER
- Filter by ‘pregnancy & birth’ life stage and required studies
- Get questions and variables
- Suggested items – similar, nearby, in and cross study

closer
Cohort & Longitudinal Studies Enhancement Resources

Why isn’t this already there?

- Started with a dream: the search platform
- Many have tried
- Limited success
- Technology supports it
- Not about technology – all about content
- Metadata not in a standard format
- Lack of managed processes for consistency

closer
Cohort & Longitudinal Studies Enhancement Resources

Content needs to be …

- Consistent
- Comparable
- Support discovery
- Re-usable
- Support processing
- Sound familiar? Metadata standards
- DDI Lifecycle

closer
Cohort & Longitudinal Studies Enhancement Resources

Why DDI-Lifecycle?

- Maintained and developed
- Mature
- International reach
- Used by major archives (DDI-Codebook)
- Fits longitudinal studies well

closer
Cohort & Longitudinal Studies Enhancement Resources

DDI is not easy

- Large and complex
- Difficult to explain to non-experts
- Even harder to demonstrate to non-experts
- Not commonly used in medical circles

closer
Cohort & Longitudinal Studies Enhancement Resources

Historical DDI is even harder

- Data goes back to 1930s
- Projects go back to 1946

- “The Schleswig-Holstein question is so complicated, only three men in Europe have ever understood it. One was Prince Albert, who is dead. The second was a German professor who became mad. I am the third and I have forgotten all about it.”
  — Lord Palmerston

closer
Cohort & Longitudinal Studies Enhancement Resources

Getting started

- Assess the scale of the problem
- Try stuff
- Don’t panic
- Assess the risks and challenges

BilbioArchives/Library Archives @Flickr

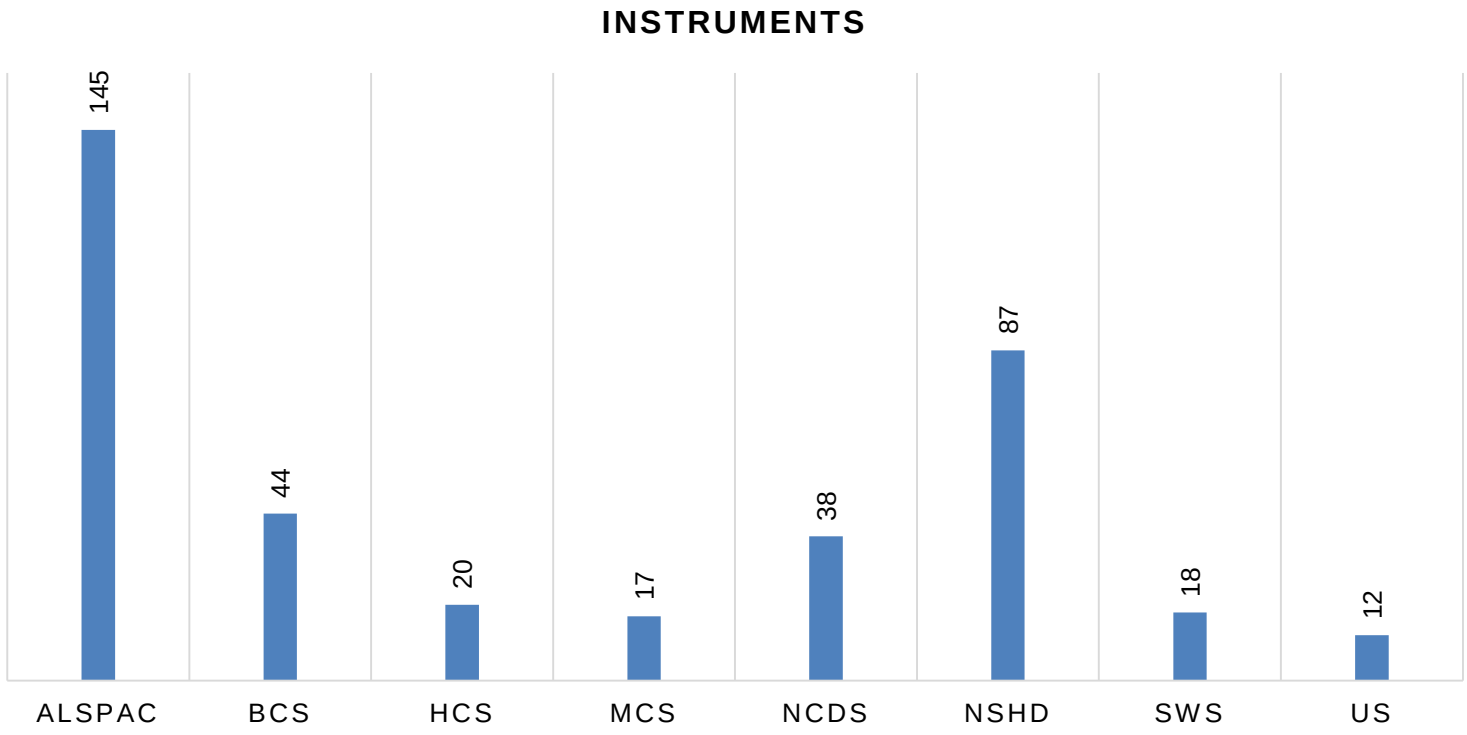
closer
Cohort & Longitudinal Studies Enhancement Resources

Metadata Project Cake Recipe

- 500g Expert knowledge
- 250g Project management
- 250g Patience (you may need to add more as required)
- 100g Software
- 500g Funding
- **1 kg of Commitment**

closer
Cohort & Longitudinal Studies Enhancement Resources

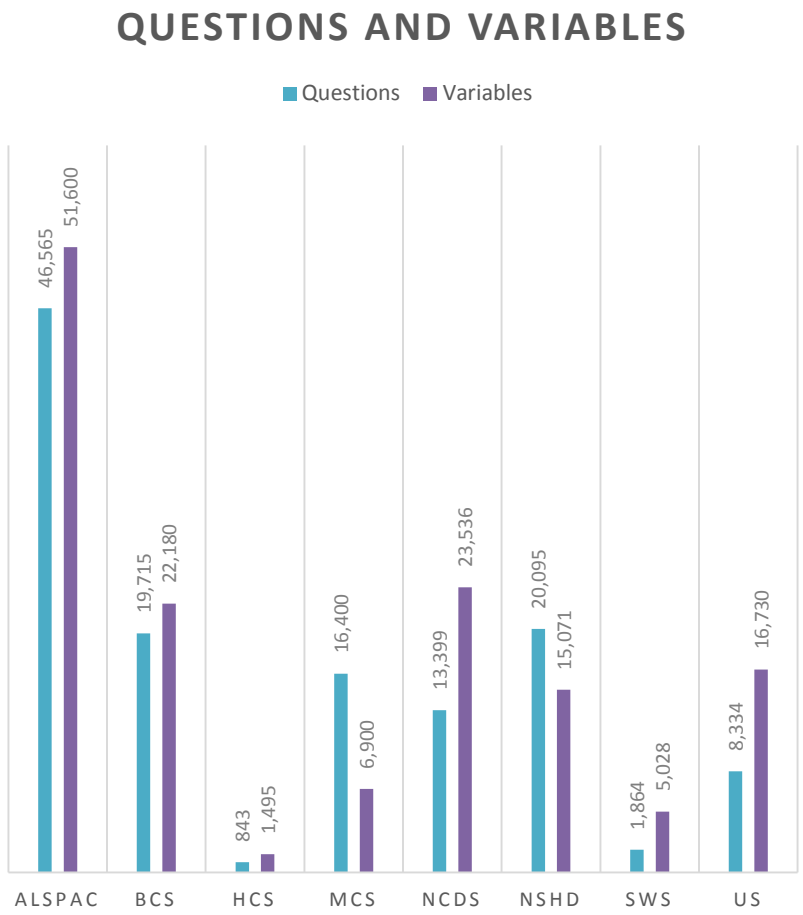
Challenges: Scale



- Instruments: 381

closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: Scale



- Huge:
  - >200k questions and variables
  - >300 questionnaires
- Ambitious – no one has attempted this scale before

closer
Cohort & Longitudinal Studies Enhancement Resources

Solution: Reduce, Reuse, Recycle

- Limited DDI profile
- Define allowable entries
- Determine content to enter
- Import from older DDI
- Import from databases, spreadsheets, etc.

closer
Cohort & Longitudinal Studies Enhancement Resources

Solution: Manageable chunks

- Unit of measurement: the questionnaire or the dataset
- Estimation:
  - Effort per questionnaire
  - Effort per question
  - Bonus for complexity
  - Controlled runs
  - Flexible
- Workflow

closer
Cohort & Longitudinal Studies Enhancement Resources

Pilot

- Had developed a tool for questionnaire entry (Caddies)
- Piloted this
- Outcomes:
  - Consistency
  - Study capacity (not about money)
  - Timelines

closer
Cohort & Longitudinal Studies Enhancement Resources

1. Ingest

- Manual entry
- Metadata Team
- Manual and standard operating procedures
- Covers all types:
  - Questionnaire
  - Clinical tool
  - Data extract

closer
Cohort & Longitudinal Studies Enhancement Resources

Principles

1. Maintain and do not alter the semantic meaning of the questionnaire
2. Do not correct the questionnaire
3. Only record what is contained within the questionnaire
4. Do not allow the data recorded (i.e. the variables) to inform the metadata archiving

closer
Cohort & Longitudinal Studies Enhancement Resources

Nothing is ever easy …

b) **If <u>no</u>, or no partner:**

Does this older child have (please tick):

you as the natural mother (but his/her natural father is not present)	1 →	answer (c) below and then go to (e)
your partner as the natural father (but his/her natural mother not present)	2 →	answer (d) on page 37 and then go to (e)
neither of his/her natural parents present	3 →	answer (c), (d) and (e)

closer
Cohort & Longitudinal Studies Enhancement Resources

… but its still possible

If no, or no partner:

- if-true branch:
  - **qc_D15_b**

    Does this older child have (please tick):

    - **1** you as the natural mother (but his/her natural father is not present)
    - **2** your partner as the natural father (but his/her natural mother not present)
    - **3** neither of his/her natural parents present

If you as the natural mother (but his/her natural father is not present) to question D15b or neither of his/her natural parents present to question D15b

- if-true branch:
  - **qc_D15_c**

    How often do you or your partner talk to the child's natural father about this older child?

    - **1** once a month or more
    - **2** less than once a month
    - **3** once a year or less
    - **4** never
    - **9** don't know
    - **7** natural father is dead

If your partner as the natural father (but his/her natural mother not present) to question D15b or neither of his/her natural parents present to question D15b

- if-true branch:
  - **qc_D15_d**

    How often do you or your partner talk to this older child's natural mother about the child?

    - **1** once a month or more
    - **2** less than once a month
    - **3** once a year or less
    - **4** never
    - **9** don't know
    - **7** natural mother is dead

qc_D15_e

Are your relations with this older child's other parent(s) :

- **1** generally warm and friendly
- **2** sometimes friendly
- **3** polite
- **4** distant
- **5** usually unfriendly
- **6** no relationship
- **7** child's other parent is dead

closer

Cohort & Longitudinal Studies Enhancement Resources

2. Verification

- Still within the Metadata Team
- Checks for accuracy
- Also checks for consistency

closer
Cohort & Longitudinal Studies Enhancement Resources

3. Mappings

- Map all questionnaires against variables
- E.g. q1 is mapped to variable N123
- Process also accommodates derived variables
- Topic mappings

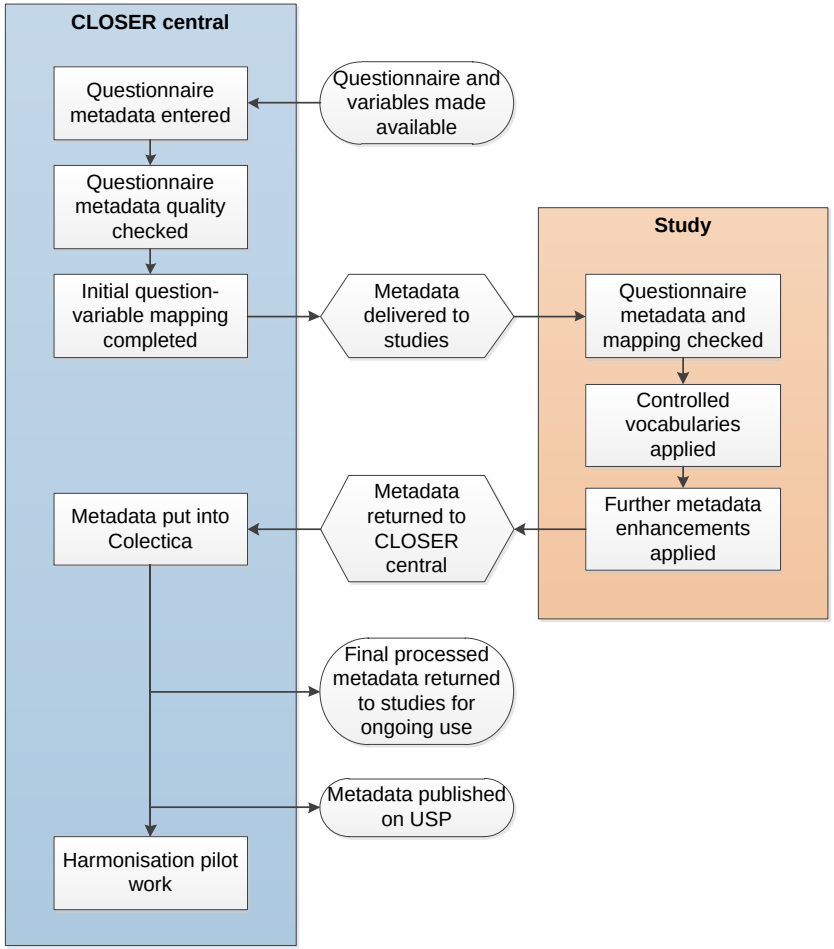
closer
Cohort & Longitudinal Studies Enhancement Resources

4. Repository ingest

- Returned to CLOSER
- Put through systematic checks before entry
- Queries from checks addressed
- Entered into Colectica repository
- Made available in USP at this point

closer
Cohort & Longitudinal Studies Enhancement Resources

Process



- Process bound – each step relies on those prior
- Efficiency/consistency of CLOSER central and subject expertise of studies

closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: varied material

- Age of questionnaires
- Condition of questionnaires and data
- Variation in methodology – PAPI, CAPI, CATI, etc.
- Medical data, images, videos, biosamples, etc.

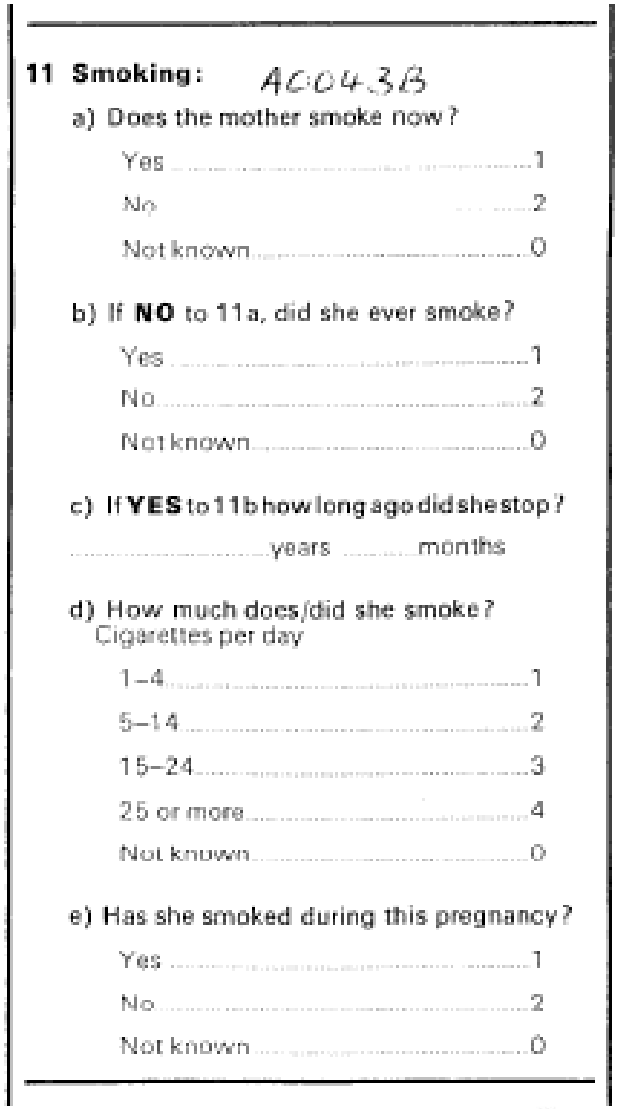
closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: Intent

- What was the intent of a question?
- Layout
- Machine logic
- Limitations of technology
- Assumptions & implied logic

closer
Cohort & Longitudinal Studies Enhancement Resources

BCS 1970



11 Smoking: AC0438

a) Does the mother smoke now?

Yes 1
No 2 (66)
Not known 0

b) If **NO** to 11a, did she ever smoke?

Yes 1
No 2 (67)
Not known 0

c) If **YES** to 11b how long ago did she stop?

..... years ..... months (68–70)

d) How much does/did she smoke?
Cigarettes per day

1–4 1
5–14 2
15–24 ..... 3 (71)
25 or more 4
Not known 0

e) Has she smoked during this pregnancy?

Yes 1
No 2 (72)
Not known 0

closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: Computers

- No machine logic
- No machine readable copies
- Limited variable creation

- Then, when computers arrive …
- MILLIONS of questions

closer
Cohort & Longitudinal Studies Enhancement Resources

MCS Birth

PDOB (PDBD, PDBM, PDBY)[3]
^And now yourself. What is your date of birth?
DATE
CHECK HH9

^And now yourself. What is your date of birth? = 'And now yourself. What is your date of birth?' if refering to first person in household (i.e. the person answering household module)
'What is *^name given at PNAM*'s date of birth?' if referring to second and subsequent members of household.

IF date of birth is don't know and person being asked about is not respondent [PDOB=DK AND PNo>1]
|
| **PAGE**
| Do you know *^name given at PNAM*'s age last birthday?
| ENTER AGE OR <Ctrl+K> FOR DON'T KNOW
| Range: 0..997
|
ELSEIF date of birth is given
|
| **PAGE** [computed using INTD and PDOB]
| Range: 0..997
|
ENDIF

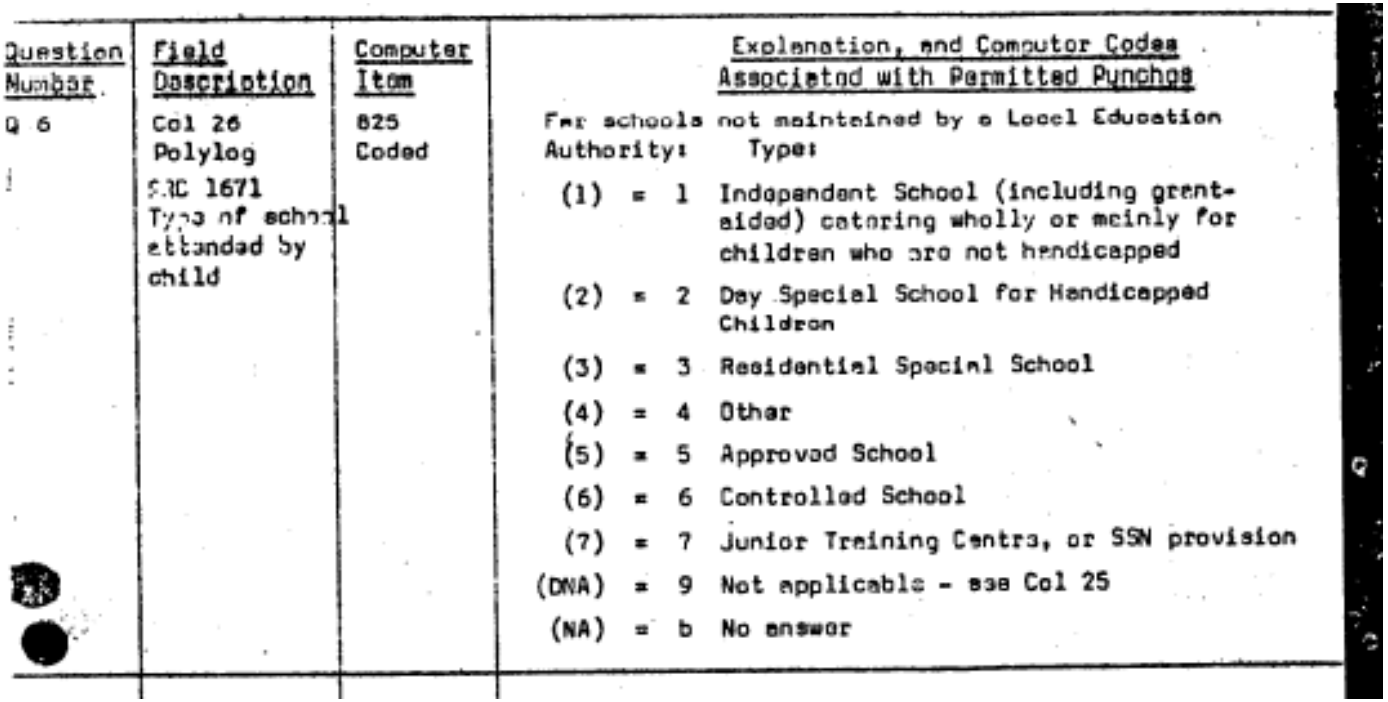
closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: Documentation

- Studies changed hands
- Formats
- What was considered important
- Changing ideas of data management

closer
Cohort & Longitudinal Studies Enhancement Resources

NCDS Guide from 1969



Question Number	Field Description	Computer Item	Explanation, and Computer Codes Associated with Permitted Punches
Q 6	Col 26 Polylog SRC 1671 Type of school attended by child	825 Coded	For schools not maintained by a Local Education Authority: Type: (1) = 1 Independent School (including grant-aided) catering wholly or mainly for children who are not handicapped (2) = 2 Day Special School for Handicapped Children (3) = 3 Residential Special School (4) = 4 Other (5) = 5 Approved School (6) = 6 Controlled School (7) = 7 Junior Training Centre, or SSN provision (DNA) = 9 Not applicable - see Col 25 (NA) = b No answer

closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: Versions

- Gender
- Country
- In-field fixes
- Attempts to increase response rate
- Change of policy

closer
Cohort & Longitudinal Studies Enhancement Resources

Challenges: consistency

- Larger the group, harder consistency is
- Agree standards
- Manual
- How long do you debate a point?

closer
Cohort & Longitudinal Studies Enhancement Resources

Extracts from the issue log

- When to use a grid?
- When is a scale not a scale?
- Images as questions
- Images as response domains
- Character set encoding
- Are corner labels on grids cosmetic or semantically significant?

closer
Cohort & Longitudinal Studies Enhancement Resources

The Madness of IDs

- What is the purpose of an ID?
- Does an ID have meaning?
- How do you ensure IDs are unique?
- Using IDs as references

- When we got to the portal interface, it all changed!

closer
Cohort & Longitudinal Studies Enhancement Resources

Solution: Communication

- Meet regularly
- Group size
  - Who needs to contribute
  - Consequences for other parts of the project

closer
Cohort & Longitudinal Studies Enhancement Resources

What happens after CLOSER?

- Documented the metadata
- Created the portal
- Made it available
- What next?
- Need sustainable processes to produce metadata
- Business as usual

closer
Cohort & Longitudinal Studies Enhancement Resources

Sustainable futures

- Flexible
  - Software independent
- Sustainable
  - Involvement with standard development
  - Survey agencies
  - Produce once, use multiple times
- Documentation and training

closer
Cohort & Longitudinal Studies Enhancement Resources

More than just metadata

- Publishing the protocols
- Publishing estimate methods
- Making CADDIES available
- Other training materials?
- Enhancing Colectica search UI
- Demonstration projects
- Future improvements

closer
Cohort & Longitudinal Studies Enhancement Resources

Good practices

- Controlled vocabularies
- Tracking software
- Methods for ensuring consistency
- Approval process
- Flexibility – one size will not fit all
- Rigidity – where you can’t be flexible

closer
Cohort & Longitudinal Studies Enhancement Resources

The integrated future

- Survey agencies
  - Make DDI a contractual requirement
  - Improved exchange of information
- Data managers
  - Use DDI as the backbone for your data
  - Metadata APIs

closer
Cohort & Longitudinal Studies Enhancement Resources