

# Lessons from a Large-Scale Metadata Archiving Team Using DDI-Lifecycle


Hayley Mills  
Metadata Officer

# Content

- CLOSER
- CLOSER Discovery
- Our set-up
- Achievements so far
- What we have learnt
- Conclusions

# What is CLOSER?

- Cohort & Longitudinal Studies Enhancement Resources
- Maximise longitudinal studies **use, value** and **impact**
- Consortium of 8 studies:

- 1931
- 
- Hertfordshire Cohort Study
  - 1946 MRC National Survey of Health and Development
  - 1958 National Child Development Study
  - 1970 British Cohort Study
  - Avon Longitudinal Study of Parents and Children (Children of the 90s)
  - Southampton Women's Survey
  - Millennium Cohort Study (Child of the New Century)
  - Understanding Society: The UK Household Longitudinal Study
- Today

# What is CLOSER Discovery?

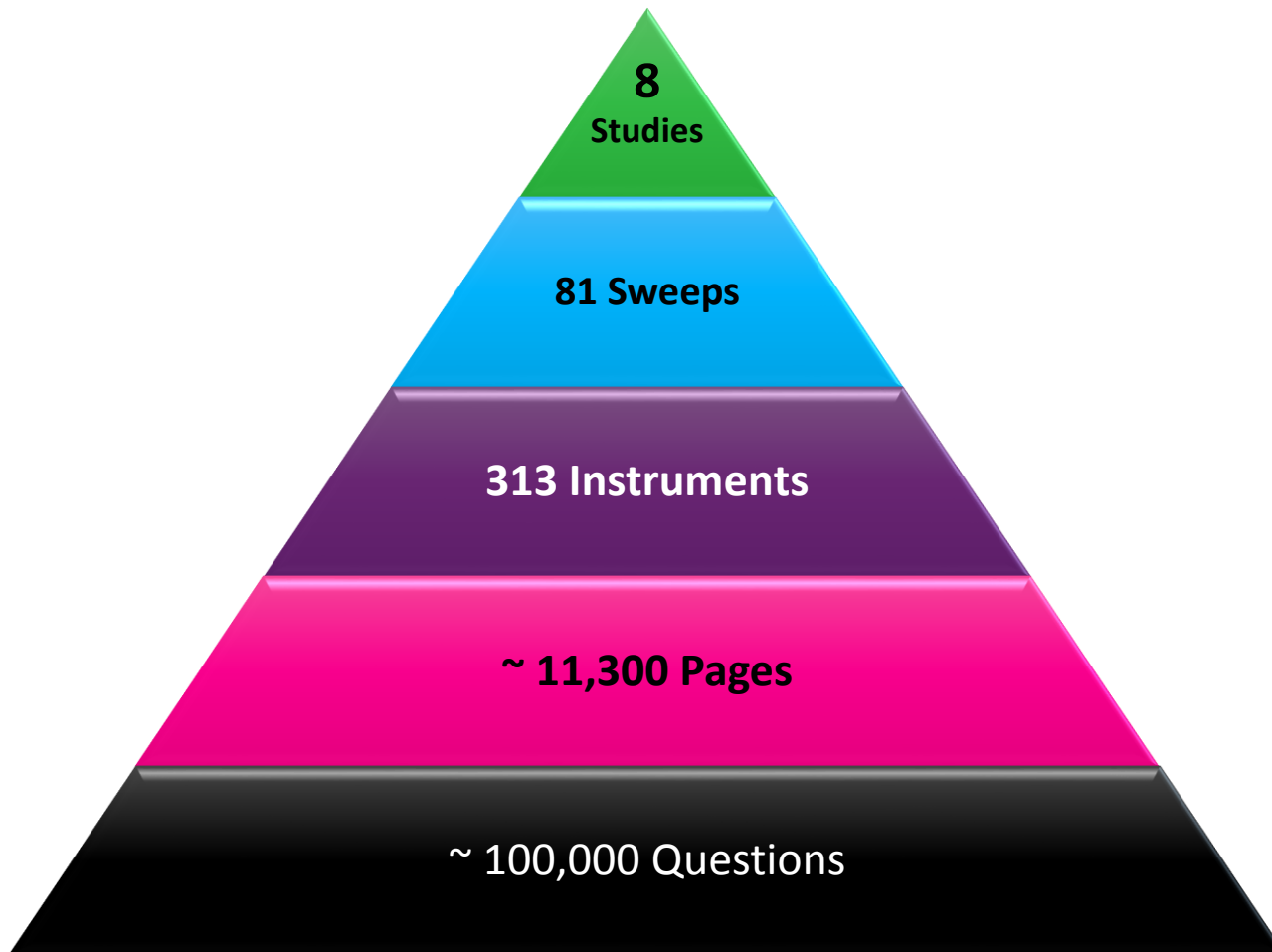
<http://discovery.closer.ac.uk/>

- Search engine for longitudinal study metadata
- Portal to variables, questions and data collection instruments
- BETA
- DDI-Lifecycle 3.2
- Questionnaire profile
  - Basic top-level Instance elements
  - Control Constructs (Sequences, Loops, Conditions, Statements and QuestionConstructs)
  - QuestionItems and QuestionGrids
  - CodeLists
  - Categories

# What is the scale?

- Massive metadata ingest programme of legacy questionnaires
- Spans 80 years from 1930s to 2012
- Very diverse range studies and instruments

# What is the scale?



# How do we get this into DDI 3.2?

- Automated software not feasible
  - Scanned paper copies
  - Not machine readable
  - Pre-computers
  - No consistency or repeatability
  - Complex

C7 Here are three things that some people of your age get up to when they are with their friends. How do you feel about each of them?

On each line you have in fit in three boxes, one for shoplifting, one for taking money by force and one for breaking into a house.

Answer a-h

This will be complete when you have put 3 ticks on each line, one for shoplifting, one for using force, and one for breaking into a house.

Shoplifting something worth less than £10

Very much Not quite a bit Not at all

Using force to get money from a stranger

Very much Not quite a bit Not at all

Breaking into someone's house to steal

Very much Not quite a bit Not at all

Tick one box on each line

Tick one box on each line

Tick one box on each line

(a) If you knew you wouldn't be caught how tempted would you be to try this?

(HC7A.1)

(HC7B.1)

(HC7C.1)

(b) How wrong do you think it would be to do this?

(HC7A.2)

(HC7B.2)

(HC7C.2)

(c) How upset would your parents be if they found out you had done this?

(HC7A.3)

(HC7B.3)

(HC7C.3)

(d) Would your friends look down on you if you had done this?

(HC7A.4)

(HC7B.4)

(HC7C.4)

(e) How likely would you be to get caught if you did this, say next Saturday?

(HC7A.5)

(HC7B.5)

(HC7C.5)

(f) How likely is it you would have to go to court if you were caught?

(HC7A.6)

(HC7B.6)

(HC7C.6)

(g) How bad do you think your punishment would be if a court found you guilty?

(HC7A.7)

(HC7B.7)

(HC7C.7)

(h) How much difference would it make to your job chances if you were caught?

(HC7A.8)

(HC7B.8)

(HC7C.8)

8. The following five statements are sometimes made about the police. For each statement about the police please say first whether you agree or disagree with the statement (give your answer in column 1). Then in columns 2, 3, 4 and 5 you are asked to say what has influenced the way you have just answered. Has it been influenced by something you've seen on TV or in the paper (tick the box in column 2), if by something which has happened to you personally, (tick the box in column 3), if by something you've been told about the police, (tick the box in column 4), or if by something else, (tick the box in column 5)?

WHAT HAS INFLUENCED YOUR OPINION ABOUT THE POLICE?

(2) Is it because of anything you've SEEN on TV/in a newspaper?

(3) Is it because of something which has HAPPENED to you?

(4) Is it because of something you have been TOLD?

(5) Is it because of some other reason?

(1) Yes, I agree No, I disagree

DO YOU AGREE WITH THESE STATEMENTS?

(a) The police in this area do their job as fairly as possible

(HC8A.1)(HC8B.1)(HC8C.1)(HC8D.1)(HC8E.1)

(b) The police are generally helpful and friendly towards young people like me

(HC8A.2)(HC8B.2)(HC8C.2)(HC8D.2)(HC8E.2)

(c) The police often mistakenly suspect young people like me of wrong doing

(HC8A.3)(HC8B.3)(HC8C.3)(HC8D.3)(HC8E.3)

(d) The police are often rough in the way they deal with young people like me

(HC8A.4)(HC8B.4)(HC8C.4)(HC8D.4)(HC8E.4)

(e) The police are always picking on young people like me

(HC8A.5)(HC8B.5)(HC8C.5)(HC8D.5)(HC8E.5)

SECTION III - Please note carefully: the information in this section is to be got from records if at all possible. If not, it may be possible, for details from mother.

PAST OBSTETRIC HISTORY - Exclude present pregnancy

21 Has the patient had any previous pregnancies (including miscarriages)?

N504

Yes ☐ No ☒

22 Was the pregnancy at a birth before, ending the pregnancy at a birth or miscarriage (the earliest first)? Record rates on two consecutive births.

Frequency Number	Date of Delivery		Sex	Birth weight	Place of Delivery	Outcome of Delivery				Complications of Pregnancy				Method of Delivery			
	Month	Year				Home	Other	Stillborn	Dead under 7 days	Stillborn	Dead under 7 days	Other complications	Other complications	Other complications	Other complications	Other complications	Other complications
1			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
2			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
3			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
4			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
5			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
6			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
7			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
8			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
9			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	
10			Y X	____lb. ____oz.	0 1	2 3 4 5	6 7 8	9	10	11	12	13	14	15	16	17	

PRESENT PREGNANCY

23 Patient's Height

(measure if not recorded, upright against the wall, without shoes)

in inches - N510

Recorded ☐ Not recorded, unable to measure ☐

Measured by midwife ☐

Not recorded, unable to measure ☐

24 Was any booking made for this delivery?

Yes ☐ No ☒

If a booking made

(a) Week original booking made?

Week ☐

(b) What kind of booking was this original one?

Domiciliary ☐

Hospital ☒

N.H.S. Maternity Home ☐

Private Nursing Home ☐

Private ward of N.H.S. Hospital ☐

Other place (specify) ☐

If original booking domiciliary

(i) Why was this booking domiciliary?

No hospital indication ☐

Hospital recommended but patient refused ☐

Hospital indicated no bed available ☐

If not for any of above reasons, specify ☐

(ii) Was this booking changed to an institutional booking during the pregnancy? If so, in which week was the change made?

Not changed ☐

Changed on the ☐ week

(iii) What was reason for change?



Cohort & Longitudinal Studies Enhancement Resources



# How do we get this into DDI 3.2?

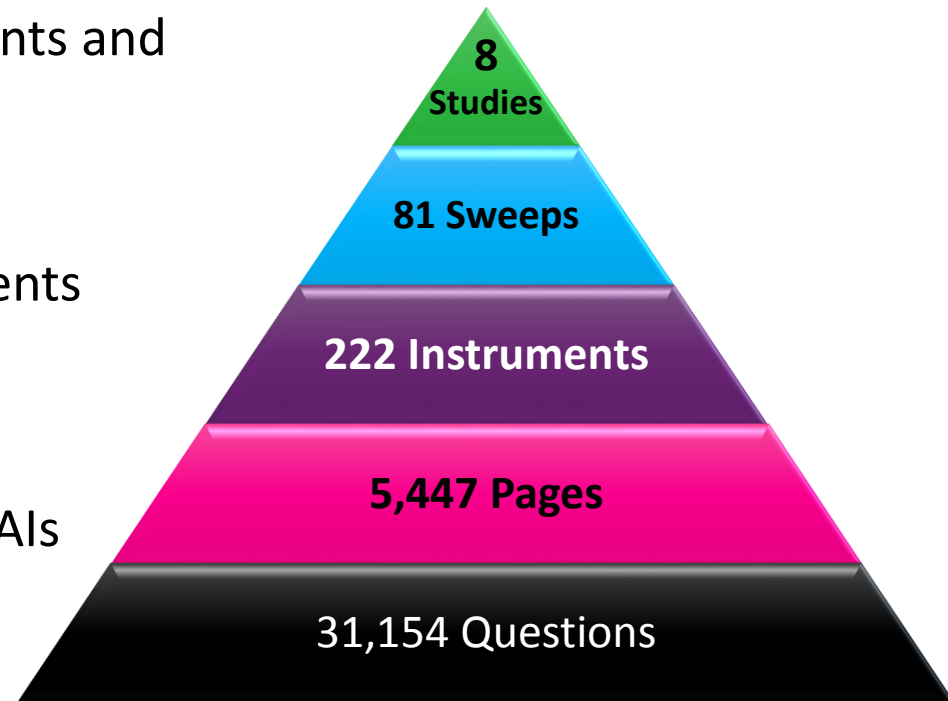
- Manual data entry
  - Consistency
  - Accuracy
  - High quality
  - Repetitive but complex data entry
- Staffing is the biggest expenditure of the project

# What is our setup?

- Metadata Assistants 2-5 and a Metadata Officer
- Centralised
  - Software limitations
  - Ensure high quality and consistency
  - Developing protocols
- Documenting software- Archivist
- Process of entry and verification

# What have we achieved so far?

- Entered over 70% of the instruments and 50% of total pages
- CAls make up 7% of total instruments but 55% of the pages
- 84% of the remaining pages are CAls



# What have we learnt?

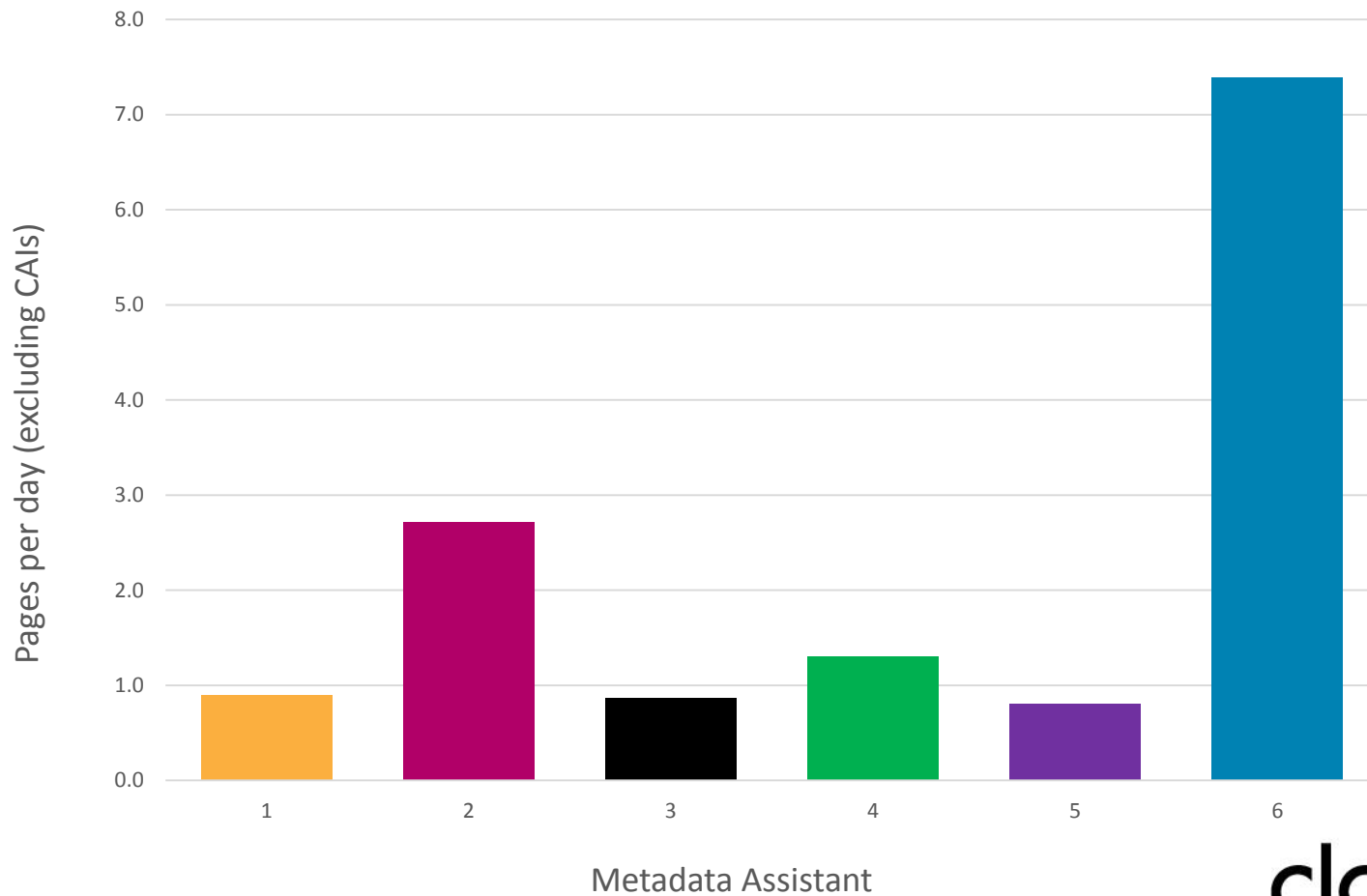
Approximate estimates

- 2.6 pages per person per day to **enter**
- 1.7 pages per person per day to **complete**
- 7 days to **enter** an average (18 pages) paper questionnaire
- ~11 days to **complete**

# What are the influencing factors?

- Team
- Tools
- Questionnaires

# Team



# Team

- Recruitment
  - Grade 5 (£20-24k)
  - Key skills
    - Good computer literacy
    - Problem solving
    - Keen on detail and accuracy
    - Motivation to develop/improve
  - Characteristics
    - Analysers and organisers (logical, fact based, organised, details, planning)
  - Motivation and retention
    - Achievement driven
    - Social science background


# Tools

- Specialist designed software – Archivist
- Simple to use and to learn
- Continuous improvements being made following feedback from the team
- Web browser so can be used anywhere
- Available for people to use <https://github.com/CLOSER-Cohorts/archivist>



# Tools

- Principles for archiving metadata
- Documentation protocols <https://wiki.ucl.ac.uk/display/CLOS/CLOSER>

 CLOSER


PAGE TREE

- > CLOSER Discovery
- > Controlled Vocabularies
- > Software
- > Standards
- > How to
  - Document a Questionnaire
    - Workflow
    - Principles
  - Archivist Elements
    - Constructs
    - Questions
    - Response Domains
    - Code Lists
  - Construct a Label
  - Step-by-Step
  - Document a Dataset
  - Use Loader
  - Map Questions and Variables
  - Apply Topics
  - Embed CLOSER Discovery's Search

Pages / ... / Document a Questionnaire @

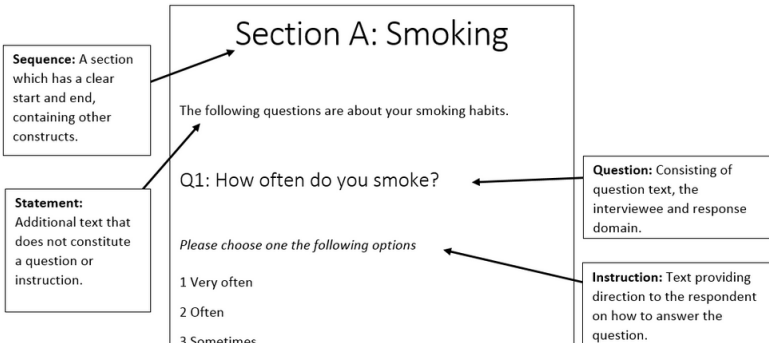
## Archivist Elements

Created by Gemma Seabrook, last modified by Hayley Mills on Aug 19, 2016

 This page is still under construction.

The various elements of Archivist are all based on the DDI Lifecycle standard. The following pages cover each of the elements of Archivist in detail. Following the description, you will be able to see case studies and further examples to help understand how complex decisions were made. The questionnaires that form a part of CLOSER are varied and complex and different ways of addressing each of the issues were developed throughout the process. The examples display a direct comparison between an actual question taken from a questionnaire, and how it translates into Archivist.

A questionnaire comprises of questions, answers and constructs which allow the respondent to navigate through the questionnaire. These processes are captured in four main elements: **Constructs**, **Questions**, **Response Domains** and **Code Lists**. The Archivist Elements are used to break down and classify metadata within a questionnaire; which can then be input accordingly. The following diagram is a basic example of a question being split into some of the Archivist Elements.



**Sequence:** A section which has a clear start and end, containing other constructs.

**Statement:** Additional text that does not constitute a question or instruction.

**Section A: Smoking**

The following questions are about your smoking habits.

Q1: How often do you smoke?

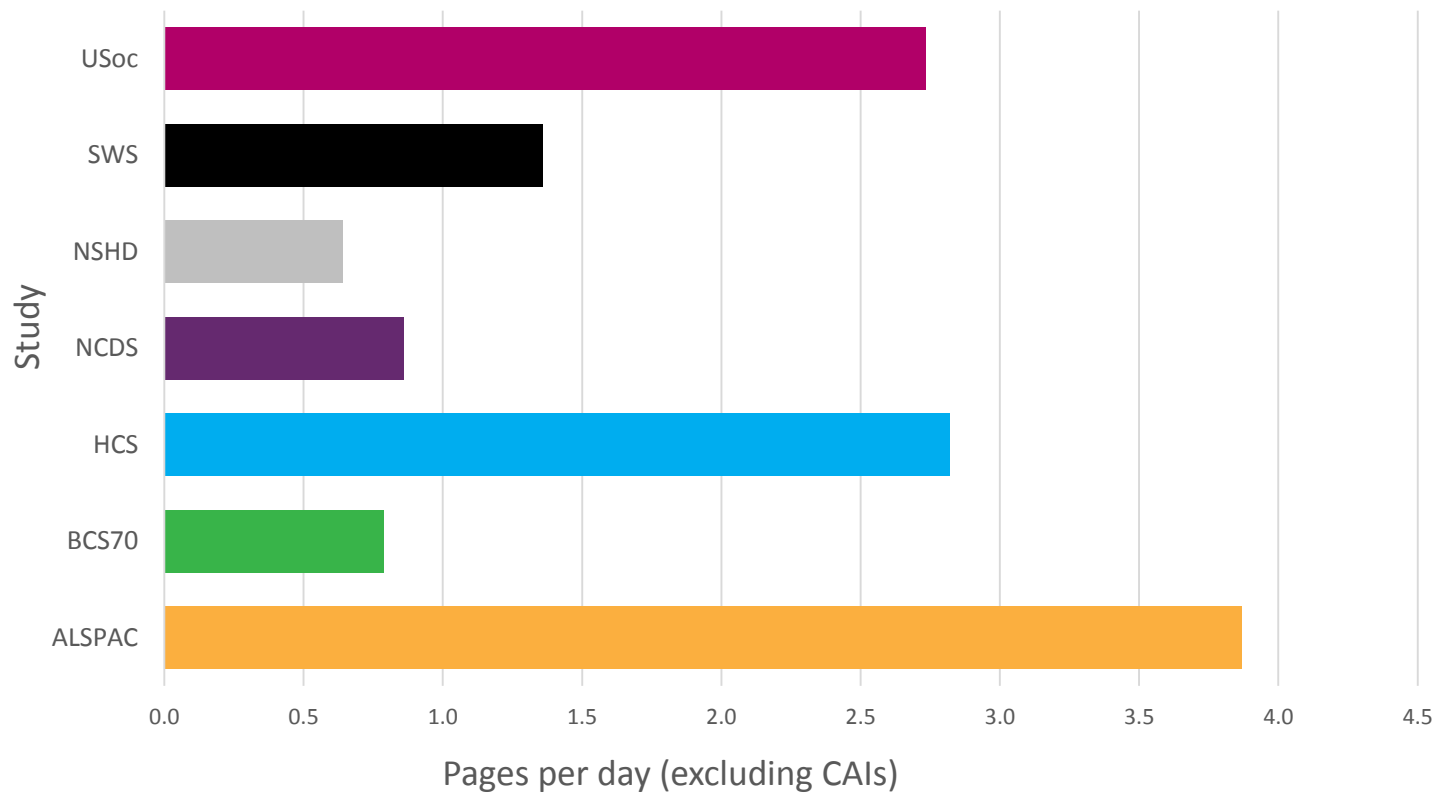
Please choose one the following options

1 Very often  
2 Often  
3 Sometimes

**Question:** Consisting of question text, the interviewee and response domain.

**Instruction:** Text providing direction to the respondent on how to answer the question.

# Questionnaires



## Q1

Please write in your date of birth:

		1	9	
--	--	---	---	--

Day Month Year

SCDOBD	SCDOBM	SCDOBY4
--------	--------	---------

## Q2

Are you male or female?

SCSEX

Male ☐

Female ☐

The first questions are about how you have been feeling recently.

Have you recently...

## Q3

...been able to concentrate on whatever you're doing?

Better than usual ☐

Same as usual ☐

Less than usual ☐

Much less than usual ☐

SCGHQA

## Q4

...lost much sleep over worry?

Not at all ☐

No more than usual ☐

Rather more than usual ☐

Much more than usual ☐

SCGHQB

## Q5

...felt that you were playing a useful part in things?

More so than usual ☐

Same as usual ☐

Less so than usual ☐

Much less than usual ☐

SCGHQC

IF SELF-EMPLOYED GO TO Q.10  
IF EMPLOYED ASK Q.4 - Q.9

Q.4 Have you been promoted while you have been with your present employer? N4245

Yes ----- 1  
No ----- 2

Q.5 Have you had any training of any kind while working for your present employer? N4246

Yes ----- 1 GO TO Q.7  
No ----- 2 ASK Q.6

Q.6 Are there any opportunities for getting training or qualifications for people doing the same sort of work as you? N4247

Yes ----- 1 GO TO Q.9  
No ----- 2 GO TO Q.9  
Don't know ----- 8

Q.7 Was this training just showing you what the job was when you first started or was it more than this? N4248

Just what the job was when started ----- 1 GO TO Q.9  
More than this ----- 2 ASK Q.8  
Don't know/Can't remember ----- 8

Q.8 Did you go on a training course either at a college or a training centre? This could include a training centre at your place of work. N4249

Yes ----- 1  
No ----- 2

Q.9 Are your wages or salary or conditions of service negotiated by a Trade Union or Staff Association? N4250

Yes ----- 1  
No ----- 2 GO TO Q.16  
Don't know ----- 8

ASK ALL SELF-EMPLOYED-OTHERS GO TO Q.16

Q.10 Have you had any training of any kind in this job? N4251

Yes ----- 1  
No ----- 2  
Don't know/Can't remember ----- 8

PC 02

Q.11 Can I just check, does your business have assets, such as property, machinery, vehicles, stocks or materials? N4252

Yes ----- 1 ASK Q.12  
No ----- 2 GO TO Q.13

Q.12 If you were to sell your business as a going concern, how much do you think you would get for it? N4253

NEAREST £ (53)(54)(55)(56)(57)  
£100,000 or more 99996  
Refused ----- 99998  
Don't know ----- 99997

Q.13 Do you pay a National Insurance Contribution? N4258

Yes ----- 1 ASK Q.14  
No ----- 2 GO TO Q.15  
Don't know ----- 8

Q.14 Do you pay just the flat rate Class 2 contribution or do you also pay a profits related Class 4 contribution? N4254

Flat rate (Class 2) only ----- 1  
Flat rate and profits related (Class 4) ----- 2  
Don't know ----- 8

Q.15 Do you receive an income on a regular basis from this work? N4260

Yes ----- 1 ASK Q.16  
No ----- 2 GO TO Q.19  
Don't know ----- 8

ASK ALL EMPLOYEES AND SELF-EMPLOYED WITH REGULAR INCOME

Q.16 I would now like to ask you some questions about income from work. On the last occasion you were paid was the amount you received - that is your take home pay - the amount you usually receive? N4261

Yes ----- 1 ASK Q.17 VERSION A  
ASK Q.18 VERSION A  
No, usually different ----- 2 ASK Q.17 VERSION B  
ASK Q.18 VERSION B  
Don't know ----- 8



Cohort & Longitudinal Studies  
Enhancement Resources

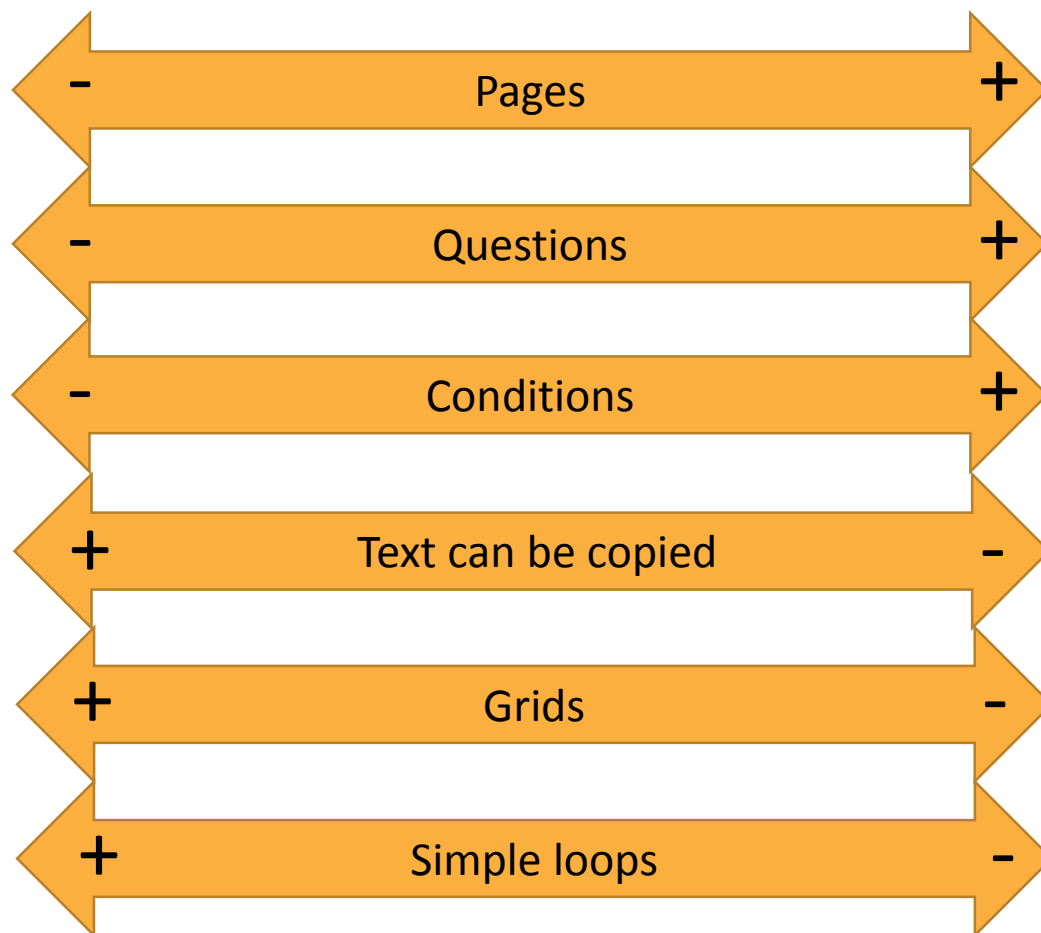
# Questionnaires

	Correlation	P value
Days vs Pages	0.17	0.03
Days vs Questions	0.26	0.001
Days vs Constructs	0.28	0.0004
Days vs Conditions	0.31	0.00009

# Questionnaires

Faster entry

Slower entry



# Conclusions

- Manual entry can produce high quality and consistent metadata to the DDI 3.2 standard on a large scale
- Staff time is the biggest expenditure so invest in the right people and tools
- Labour intensive but this standard will be expected in the future and it will increase the discoverability of the data
- Questionnaire design significantly affects speed of entry and varies greatly between studies

# Conclusions

- Don't put off entering legacy or paper questionnaires as the problem will only increase
- Invest in documenting questionnaires early in the process to future proof your data use
- Software and documentation is available to use and share

# Thank you