Protocol Development for Large-Scale Metadata Archiving using DDI-Lifecycle

by William Poynter[1] and Jennifer Spiegel[2]

Abstract

A substantial part of CLOSER (Cohorts and Longitudinal Studies Enhancement Resources, www.closer.ac.uk) is to electronically document over 100,000 questions from 379 instruments for use in a metadata discovery platform (MDP). The questionnaire profile of DDI-Lifecycle 3.2 is the specification used by CLOSER. During the metadata ingest process, a set of four entry principles were developed to outline and protect the overall standard by which the metadata is being recorded. To supplement the four principles, a protocol manual is being organically developed, including real examples of procedure. However, there are situations where the entry principles conflict with each other and DDI-Lifecycle, therefore one or more must be broken in order to adhere to the more significant principles.

This paper also discusses how the selected DDI profile (the use of Question Grids and Interview Instructions) and the tools used for ingest (CADDIES) affect the principles and methodology used to document the metadata.

Keywords: archiving, large-scale, principles, integrity.

Introduction

CLOSER (Cohorts and Longitudinal Studies Enhancement Resources, www.closer.ac.uk) is a five year project that aims to bring together nine of the UK's longitudinal cohort studies, by producing a metadata discovery platform (MDP).

A key element of the CLOSER project is to archive the historic questionnaire metadata for the nine studies involved. DDI-Lifecycle was the obvious choice as the standard with which to store the metadata that powers the MDP..

Materials and Methods

To ingest the vast quantity of metadata within a useful timeframe, multiple data entry specialists are required. Therefore, to provide a consistent approach and quality to the metadata entered extensive training is also required. The full set of DDI-Lifecycle items is deemed

Study	Questionnaires	Questions	Variables
ALSPAC	143	30033	52285
BCS	44	18235	22180
HCS	20	900	1570
MCS	17	16400	6900
NCDS	38	10246	22750
NSHD	87	19419	20000
SWS	18	1362	2028
US	12	8334	
Total	379	104929	>127713

Table 1: Eight of the nine studies within the CLOSER project, showing their estimated counts of questionnaire, question and variable. The ninth study, Life Study, is omitted from this

too large to effectively train multiple members of staff to enter metadata to the same standard, thus a subset of the DDI-Lifecycle profile is used. (ddiadmin, 2014)

Entry Principles

The Questionnaire Profile provides the optimum level of detail to preserve as much questionnaire metadata and metadata integrity as possible. (ddiadmin, 2014) To best ingest metadata to the standard set by the selected profile, it is effective to use a custom made tool. The CLOSER project developed and maintains a metadata editing tool dubbed CADDIES (CLS Abridged DDI Editor for Surveys). (Gierl, n.d.)

CADDIES is an open-source[3] GUI-based tool that allows the user to edit metadata with little or no understanding of DDI-Lifecycle. CADDIES also provides the user with basic assistance by validating required fields to maintain valid DDI.

CLOSER documents metadata using an atypical approach; metadata is recorded from the questionnaires and not from the datasets collected as in many other similar projects. (Curran, et al., 2013) (Anon., 2014) Before beginning to enter metadata it is a practical necessity to construct a set of principles by which the entry process will follow to create a consistent record. CLOSER adheres to the following principles during metadata entry, in descending order of significance.

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

The first principle listed above is the key objective of the CLOSER ingest project, hence it is most significant and must not be broken or ignored under any circumstances. The following three principles are derived practices in order to follow the first principle as consistently as possible. The order of significance reflects the potential of losing semantic meaning by breaking the principle. For example, breaking principle four is less likely to have an effect on the semantic meaning, than breaking principle three.
In addition to the four principles listed above, a metadata entry manual is created to comprehensively describe the input decisions taken and how they are resolved. Therefore the manual grows organically over time during the process of entering different questionnaires, from different studies, conducted over different time periods (1946-2010).

Question Grids

Within the chosen DDI profile there are two items used to document a question, **QuestionItem** and **QuestionGrid**. **QuestionItem** is used to document most simple and standard questions, but in accordance with DDI-Lifecycle 3.2 CLOSER is using **QuestionGrid** to document complex question formats. The use of **QuestionGrid** is crucial to both: reduce the quantity of separate DDI entities used to document a single complex question,

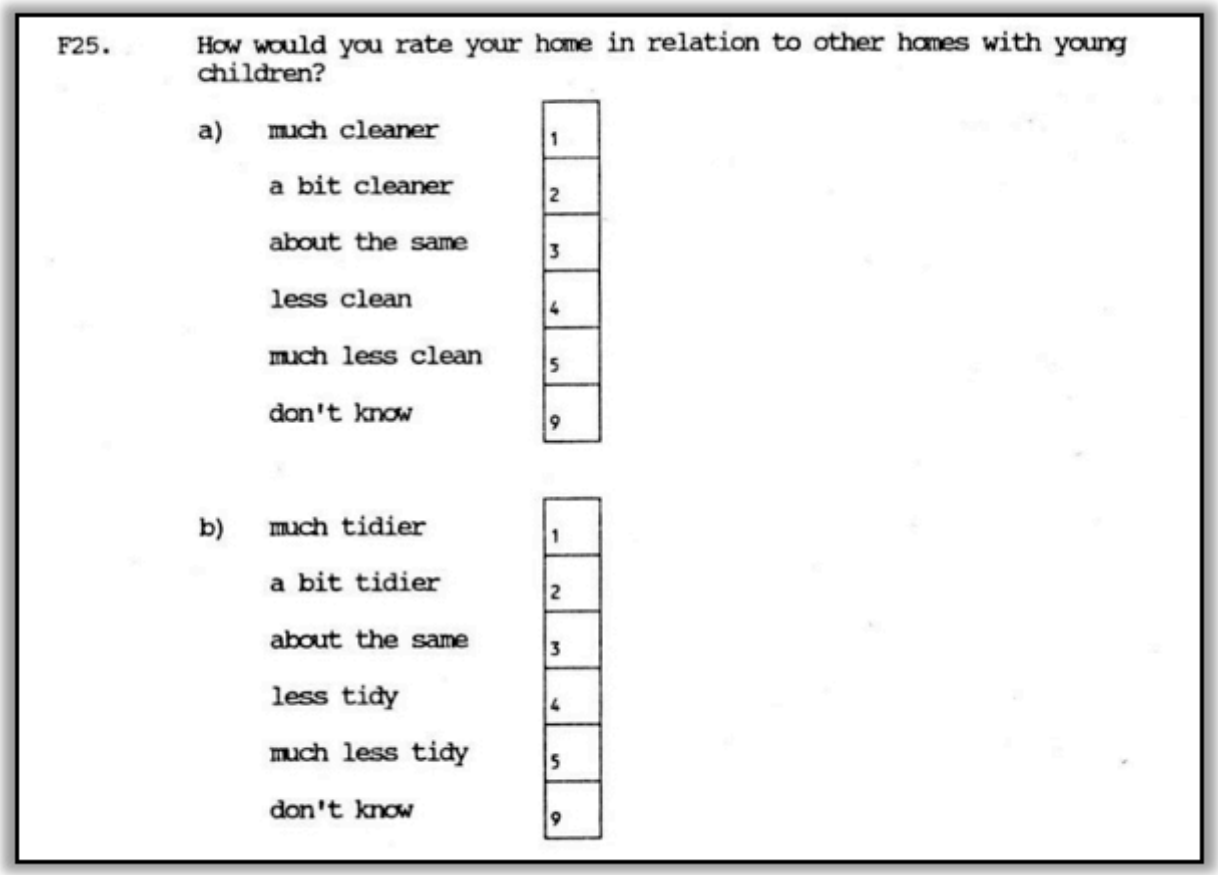


Figure 1: An example taken from an ALSPAC questionnaire depicting only one question literal, but two separate opportunities to respond.

and to better relate an overarching question context without the use of repetition.

Results and Discussion

During metadata entry it is instantly apparent that the irregularities within questionnaire design stress and strain the seemingly simple entry principles. Therefore the order of significance applied to the entry principles is used to inform a consist approach that also preserves the integrity and as much of the original meaning as possible.

Fourth Principle – Don't allow the variables to inform the entry

The fourth principle is the least significant principle. In the spirit of creating a product with maximum usefulness it is clear that on certain occasions it is advantageous to consider the variables produced for the questionnaire to educate a decision on how best to progress. For example, the question(s) shown below (Figure 1) has a single question text, but two separate code answers. In this situation two methods of documentation present themselves:

1 copy the question text and produce two QuestionItems, each including one of the two code answers,
2 do not copy the question text, but rather create a single QuestionItem, with two response domains.

Both techniques of documentation have pros and cons, but once the fact that there are two data variables has been considered, it becomes simpler and more direct to use the first method of documentation. This conclusion flouts the fourth entry principle, but produces the best documentation of this irregular question format.

It should be noted that, although there are examples of where it is beneficial to break the fourth principle, it is still import to keep to the principle in the vast majority of situations.

Third Principle – Do not add to the metadata

It may be thought that keeping to this principle is exceptionally straightforward, but a real-life application of metadata archiving shows that this is not always the case. The predictable situation of missing words, letters or punctuation with literal strings presents no issue at all, as the incorrect string can be documented as is and any mistakes can be tagged in order to improve the search functionality. For example the question text,

"What is your mther's age?",

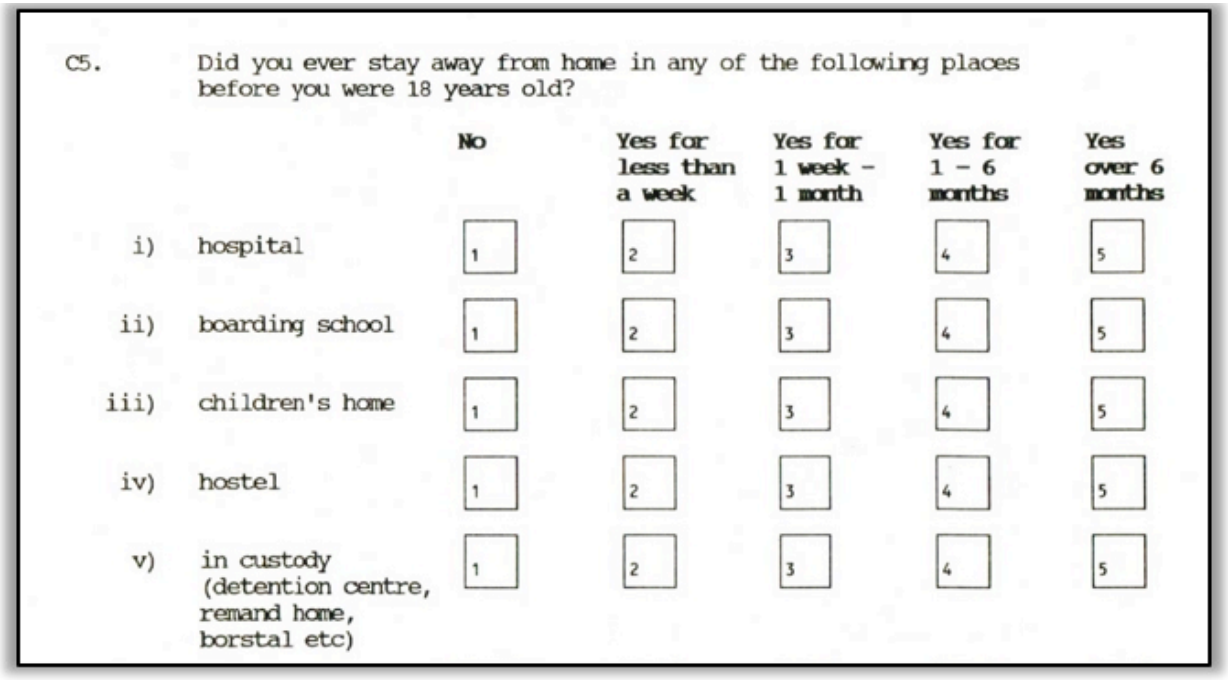
is documented as-is, but the word "mther" is tagged to be included in searches performed using the word "mother".

However, an issue with adhering to this principle arises when there is no text at all. The two most common examples of this are:

1. **QuestionGrids** with missing headings
2. **ConditionalConstructs** with no text

In the question shown below, it is fastest and most robust to document this question as a single columned **QuestionGrid**, but this creates the issue that there is no heading for the column. As DDI requires text for all column labels, it is CLOSER's protocol to use a hyphen to represent the missing text, and therefore a character has been added that was never in the original questionnaire. CLOSER selected a hyphen to represent all missing column labels, because the hyphen is a visible character making it quick and easy to identify. If it is discovered that a column header exists that is genuinely just a hyphen, it is trivial to update CADDIES to use a different character to represent the missing column header (e.g. '=' or '_').

As previously mentioned, the second common issue with the third principle is caused when the questionnaire does not provide any text to indicate a condition. Typically a textless condition is



C5. Did you ever stay away from home in any of the following places before you were 18 years old?

		No	Yes for less than a week	Yes for 1 week - 1 month	Yes for 1 - 6 months	Yes over 6 months
i)	hospital	1	2	3	4	5
ii)	boarding school	1	2	3	4	5
iii)	children's home	1	2	3	4	5
iv)	hostel	1	2	3	4	5
v)	in custody (detention centre, remand home, borstal etc)	1	2	3	4	5

Figure 2: An example taken from an ALSPAC questionnaire depicting a potentially useful situation to use QuestionGrid, but there is no column label.

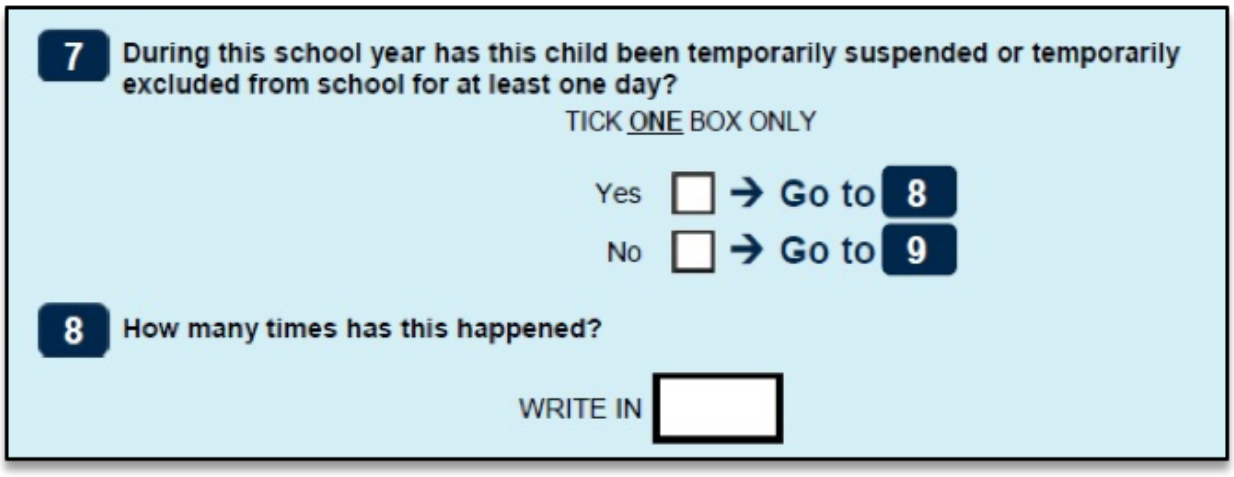


Figure 3: An example taken from an MCS questionnaire demonstrating the practice of indicating a condition using an arrow and a navigational statement 'Go to ...'.

denoted by an arrow, which is contextually meaningless and difficult to ingest into the selected DDI profile as a logic flow indicator. (see Figure 3)

As a blank condition has no use and is not permitted within DDI, it is necessary to create a string that accurately describes what is characteristically shown by the arrow. The invented conditional string to solve the example above is

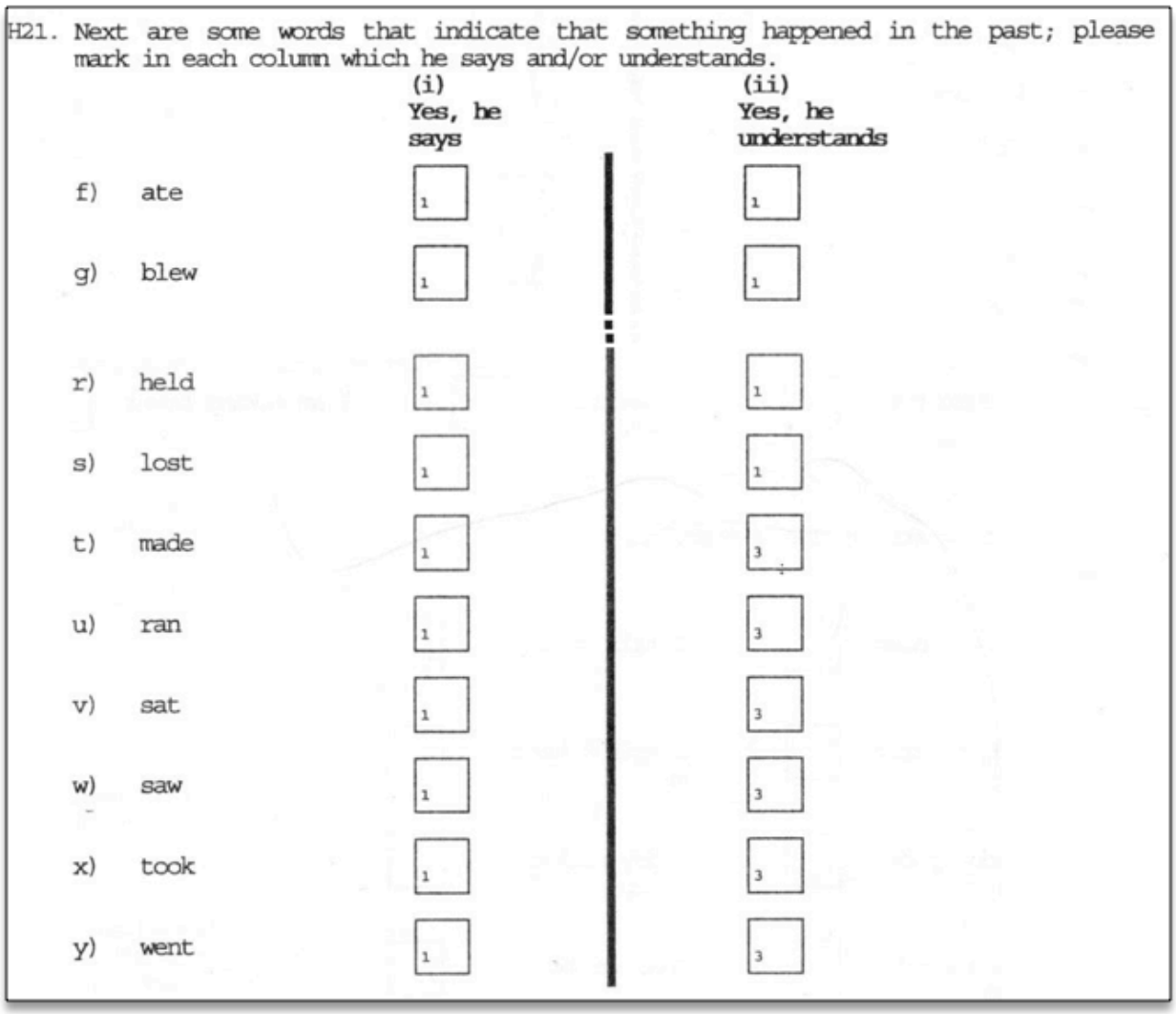


Figure 4: An example taken from an ALSPAC questionnaire displaying a perceived error in coding the second column.

"If Yes to question 7".

The formula for the invented condition text is

"If " + [category] + " to question " + [question number].

The invented condition text is in obvious disagreement within the third principle, but without inventing text for a condition, a condition cannot be recorded and therefore metadata regarding the flow of the questionnaire will be lost.

Second Principle – Do not correct the metadata

As with the third and fourth principles, there are issues with trying to follow the second principle too rigidly, although it should be mentioned that these issues are rare and are therefore dealt with on a case-by-case basis.

Predominantly, it is straightforward to handle mistakes within a questionnaire as shown in section 3.2 above, but there are a few rare cases where documenting the metadata in its original condition will directly inhibit the usefulness of the documentation process.

As can be seen in Figure 4 the code values change partway down the right hand column. This causes an issue while documenting as it is not possible to change a CodeAnswer part way down a QuestionGrid, therefore the question can no longer be documented as a single QuestionGrid. In addition to this, it is highly unlikely that the change of code value was intentional; so it is also highly unlikely that the dataset would reflect the mistake. In this situation it is deemed appropriate to violate the fourth principle and check if the dataset contains any values of 3, or were they all recorded as 1. If it is confirmed that the code values of 3 are a mistake then correcting the code values provides a more accurate documentation of the structure of the question and how the question is mapped to the data collected.

First Principle – Maintain the semantic meaning

As mentioned previously this principle is the most significant and therefore is protected most vigorously. Using the selected DDI profile it is possible to follow this principle either without issue or by breaking one or several of the other principles.

While documenting CAI (computed-assisted interview) questionnaires, to follow the first principle it is necessary to use InterviewerInstruction, which is not contained within the original DDI profile selected by CLOSER.

Interviewer Instruction

InterviewerInstruction is an integral part of documenting CAI questionnaires due to the fact that important details of the

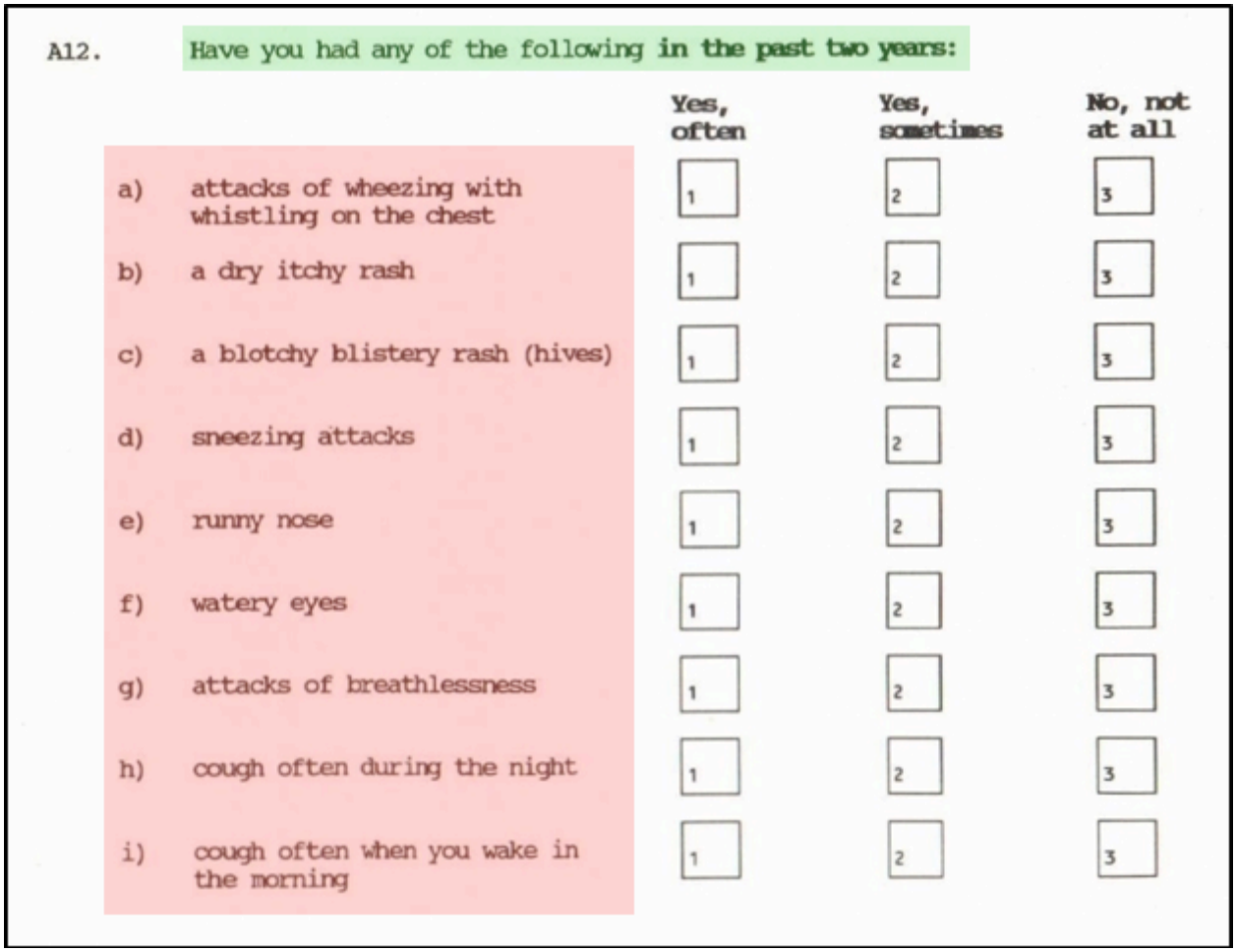


Figure 5: An example from an ALSPAC questionnaire highlighting the separate components within a complex question.

question are frequently not contained within the question text, but as instructions instead. Using a statement to hold the instructions is an ineffective approach as important question information may become separated from the QuestionItem. It would also violate the first principle to record the instructions within the question literal, as this would suggest the interviewee received the instruction text directly.

When applying the revised DDI profile to paper questionnaires it becomes apparent that there is an uncertainty as to whether InterviewerInstruction applies to instructions to the interviewer or from the interviewer. Therefore InterviewerInstruction is also used to document instructions about how to fill in a question within paper questionnaires. For example

"Tick only one box",

is documented as an InterviewerInstruction.
For the tradeoff of using a slightly more complex DDI profile, the documentation of CAI questionnaires is made much more accurate and the workload is reduced for documenting paper questionnaires, because InterviewerInstruction is reusable.

Question Grid

CLOSER uses question grids in two fundamentally different topologies. The first is the typical application, which documents complex questions with multiple dimensions of size greater than 1, which form the grid axes. (Thomas, et al., 2014) The second is a less ordinary use, to document questions constructed from a principal string, followed by multiple secondary strings. (see Figure 5)

The question above has three key methods by which it could be documented:

1. The principal text (green) is recorded as a sequence with each secondary text (red) being recorded as its own **QuestionItem**
2. The principal text (green) is concatenated to each secondary text (red) to be recorded as nine separate **QuestionItems**
3. Using a question grid, the principal text (green) becomes the question text, while the secondary text (red) becomes a code list that forms the y-axis of the grid

CLOSER uses the third technique to document this question. The first technique, involving a sequence, is not used because principal text (green) is too important to the context of the secondary texts to be loosely associated using a sequence. Although the second technique preserves as much context as the third technique, it is less preferable due to the increased entry overhead (creating nine **QuestionItems** instead of one **QuestionGrid**) and the redundancy caused by copying the principal text eight times.

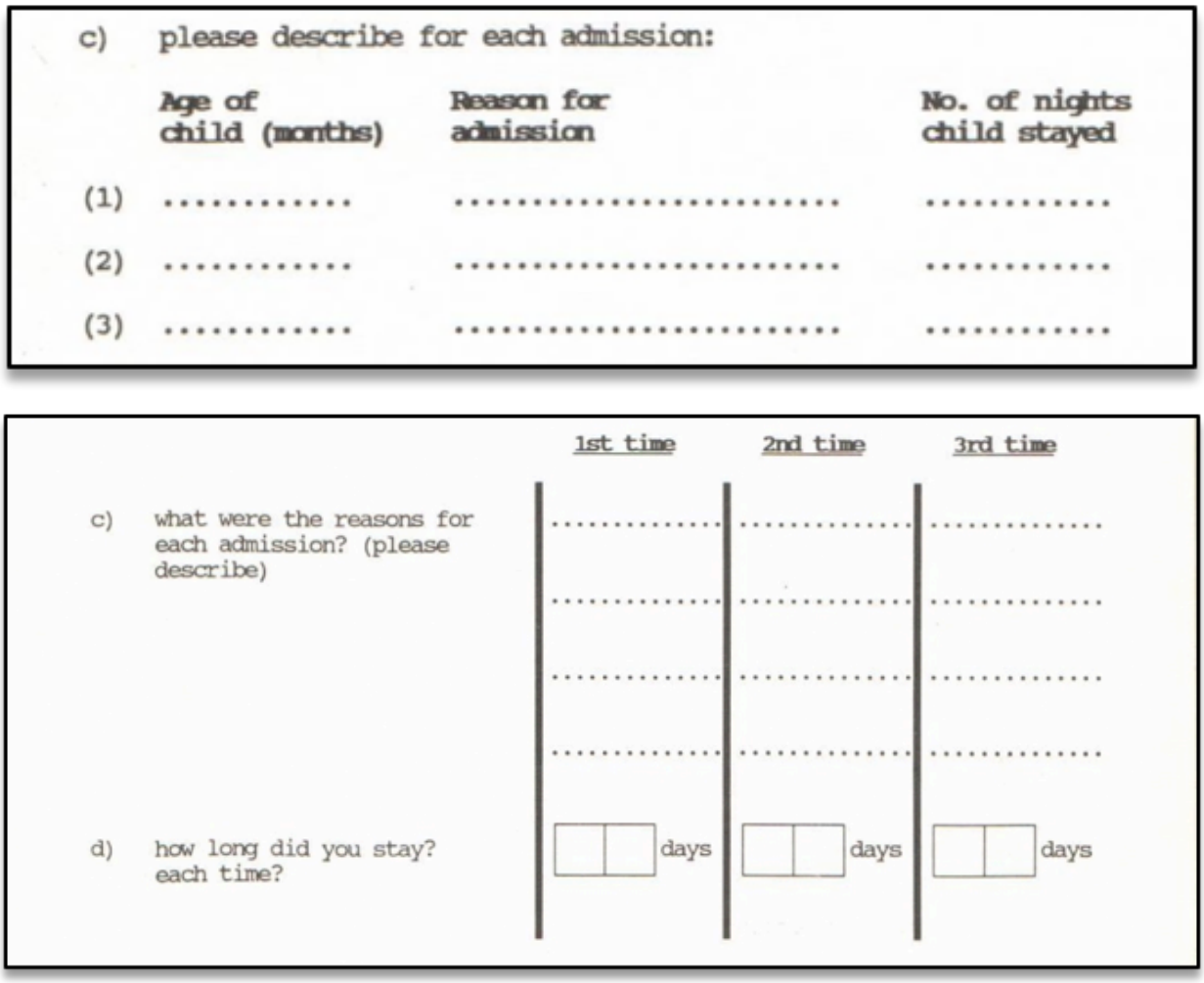


Figure 6 & 7: Two examples taken from ALSPAC questionnaires. Their different documentation topologies arise because Fig. 6 has principal text, while Fig. 7 does not.

Question Grids vs Question Loops

Although question loops (Loop) are normally used to document loops within CAI questionnaires, there are situations where it is sensible to use question loops to document complex questions within paper questionnaires. Shown in Figure 6 and 7 are two similar question topologies that require documenting using two entirely different methods.

The first is a textbook example of a QuestionGrid, but the second cannot be documented using a QuestionGrid as there is no overarching question text. Therefore in the second example the two questions have to be documented as two QuestionItems, contained within a loop from 1 to 3.

Conclusion and Future Work

The most significant conclusion that can be drawn from CLOSER's progress is that any large-scale ingest process has to be organic. It is impractical to attempt to develop rules that will meet all of the vastly different instruments within a large-scale ingest, before the process has begun. Therefore starting with principles that will guide the creation, adaptation and interpretation of rules, is far more effective and creates a more consistent result.

Using the entry principles outlined above, structure can be given to how to approach each situation consistently. Each time a principle has to be broken, the exact parameters for why it must be broken should be recorded along with exactly how to break it. The result of this process will be a documented standardised set of practices that ultimately can be repeatable and comparable.

Assuming that the overall goal of any future metadata ingest is to preserve the semantic meaning, then the principles outlined here should be applicable regardless of technology, time-period or instrument type.

The use of **QuestionGrid** to record questions with two dimensions of length greater than 1 allows for much neater and easier metadata entry. Also, there are fewer opportunities for an error to be added to the metadata, due to fewer constructs being used. However, it is currently unknown whether using **QuestionGrids** to record simpler questions (i.e. questions with only one column) will be beneficial. Any advantages in the entry process, while entering single-column **QuestionGrids**, are subject to the tool used for entry; therefore a final verdict on this approach can only be reached once the DDI output has been used.

Further ingesting work will be performed by CLOSER, allowing the entry principles to be further tested and developed from a greater sample of instruments. Work by other organisations could also be conducted to test the consistency and integrity of metadata entry both with and without the entry principles, using two different teams of people.

Acknowledgements

CLOSER is funded by the Economic and Social Research Council (ESRC) and the Medical Research Council (MRC). It has been awarded a core grant of approximately £5 million for 2012 to 2017. The funding was made possible by a landmark contribution from the Government's Large Facilities.

We would also like to mention the significant contributions and support provided by Claude Gierl, Jon Johnson and Gemma Seabrook.

References

Anon., 2014. The RAND HRS Data (Version N). [Online]
Available at: http://hrsonline.isr.umich.edu/modules/meta/rand/randhrsn/rnd_Ndd.pdf
[Accessed 17 11 2014].
Curran, P. et al., 2013. The Medical Research Council Gateway. [Online]
Available at: http://www.eddi-conferences.eu/ocs/index.php/eddi/EDDI13/paper/view/62
[Accessed 17 11 2014].
ddiadmin, 2014. DDI Alliance. [Online]
Available at: http://www.ddialliance.org/survey-metadata-reusability-and-exchange
[Accessed 17 11 2014].
Gierl, C. & Jon, J., 2013. How Do We Manage Complex Questions in the Context of the Large-Scale Ingest of Legacy Paper Questionnaires into DDI-Lifecycle?. [Online]
Available at: http://dx.doi.org/
[Accessed 17 11 2014].
Gierl, C., n.d. CADDIES. [Online]
Available at: https://github.com/claude-uk/caddies
[Accessed 17 11 2014].
Thomas, W. et al., 2014. DDI Alliance. [Online]
Available at: http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/HighLevelDocumentation/DDI_Part_I_TechnicalDocument.pdf
[Accessed 28 11 2014].

Notes

1. w.poynter@ioe.ac.uk, UCL Institute of Education, London, UK
2. j.spiegel@ioe.ac.uk, UCL Institute of Education, London, UK
3. CADDIES is currently available on GitHub.com as an open-source alpha release, i.e. is not ready for external distribution and collaboration.