 DOI: http://dx.doi.org/10.3886/eDDIGierl

Proceedings of EDDI13
5th Annual European DDI User Conference
December 2013, Paris, France

How Do We Manage Complex Questions in the Context of the Large-Scale Ingest of Legacy Paper Questionnaires into DDI-Lifecycle?

Claude Gierl[1], Jon Johnson[2]

Abstract

The Centre for Longitudinal Studies (CLS) and the CLOSER (Cohorts and Longitudinal Studies Enhancement Resources) project are in the early phase of a large scale metadata ingest programme, largely from historic paper questionnaires. The ultimate goal of the project is a Unified Search Platform (USP) which will make the metadata of 9 of the UK's birth cohort studies available and searchable online.

Part of the building blocks of the project is an in-house editor used to capture the metadata. While DDI3.2 certainly seems to offer enough flexibility to handle most of the complex or tabular questions encountered in the legacy questionnaires, the finite scope of the editor and the skill level of its intended users lead us to constrain the DDI3.2 profile we intend to use.

How we design the core structure of this editor and which DDI3.2 profile we implement in order to deal with complex questions has a wide range of repercussions for the rest of the project infrastructure. Our choices will affect issues such as how far we will be able to remain sufficiently true to the original questionnaires, how efficiently the searching of the USP will operate, how easy it will be to assign concepts, mappings and comparison schemes and, more generally, the consistency of the metadata across surveys. These repercussions need to be examined in order to weigh up our preferred options.

Keywords: ddi3.2, questions, non-trivial, high-volume, historic, profile.

1 Introduction

The Centre for Longitudinal Studies (CLS) is a national resource centre which has responsibility for three of Britain's internationally renowned birth cohort studies: the 1958

[1] c.gierl@ioe.ac.uk, Institute of Education, London, United Kingdom.
[2] J.johnson@ioe.ac.uk, Institute of Education, London, United Kingdom.

National Child Development Study, the 1970 British Cohort Study and the Millennium Cohort Study. The Centre conducts regular surveys of the cohorts, analyses their data and assists the wider research community in using the data.

CLS is participating member of CLOSER (Cohorts and Longitudinal Studies Enhancement Resources, www.closer.ac.uk), a five year project which started in October 2012. This project aims to bring together nine of the UK's longitudinal and cohort studies to maximize their use, value and impact, both within the UK and abroad. One of the main outcomes of the project will be a Unified Search Platform which will allow researchers to find the variables they need for their analyses by searching across all the studies involved in the project.

This project is a one-time only opportunity to realise this type of resource and it is therefore critical that we get all the main components and architecture right first time. One of the elements in the set of building blocks necessary for this resource to achieve its aims is the appropriate management of complex questions and this is what this analysis sets out to investigate.

2 The Legacy Questionnaires

The following studies are included in the project:

- Avon Longitudinal Study of Parents and Children
- British Cohort Study (1970)
- Hertfordshire Cohort Study
- Millennium Cohort Study
- National Child Development Study (1958)
- Southampton Women's Study
- National Survey of Health and Development (1946)
- Understanding Society

Table 1 provides an indicative measure of the scale of metadata to be ingested. The sweeps are the repeat data collection phases that form part of a longitudinal study. The survey instruments are either computer-assisted interviewing (CAI) or paper questionnaires.

The challenges of the project reside not just in the size of the sets and in the harmonization requirements of heterogeneous variables across studies but also in the inconsistency of the questions between sweeps within a study.

Table 1: The CLOSER questionnaires

	ALSPAC	BCS	HCS	MCS	NCDS	SWS	NSHD	US
Sweeps	19[3]	10	2	4	10	9	37	5
Questions	30033	18235	900	16400	10246	1362	19419	8334
Variables	52285	22180	1570	6900	22750	2028	20000[4]	NK[5]
Survey inst.	119	40	19	14	31	18	86	12
- CAI	7	7		7	3			5
- Paper	112	33	19	7	28	18	86	7

3 The CADDIES Survey Editor

Figure 1 shows a sample form of the survey editor. The name, question text and intent are free text entry fields. Answers (ResponseDomain or StructuredMixedResponseDomain) are selected via a menu and added to a table.

In order to ingest the metadata of our legacy surveys, we use various software parsers for the metadata which is already in some sort of electronic format. For paper documents, we use an online editor: CADDIES (CLS Abridged DDI Editor for Surveys).

The editor was written in-house using the Ruby-on-Rails (www.rubyonrails.org) framework with the metadata held in the sqlite3 relational database (www.sqlite.org). Whereas more comprehensive DDI3 software such as Colectica are able to manage a far richer set of DDI3 elements (see Iverson and Smith, 2012), our editor is based on a direct mapping between a fixed set of DDI3 elements and a relational schema. Amin et al (2011) provide a good overview of the issues related to DDI management in relational databases.

The current version of the editor is based on DDI3.1. Its profile includes basic top-level Instance elements, Control Constructs (Sequences, Loops, Conditions, Statements and QuestionConstructs), QuestionItems, CodeSchemes and Categories. Users enter metadata from a questionnaire through web forms which correspond to specific elements or attributes of the DDI3 schema.

Users are encouraged to enter metadata from the bottom up, i.e. to start with the Categories. These are then accessible via a graphical user interface (GUI) menu when building CodeSchemes. The relational database then holds these selections as references (foreign keys). The same process of selecting via GUI menus the lower elements to be referenced occurs at any of the next levels mapping the DDI3 reference model.

[3] ALSPAC uses a continuous survey mode and does not have sweeps. We divided the set into arbitrary sweeps for easier management.

[4] Approximate value

[5] Not known

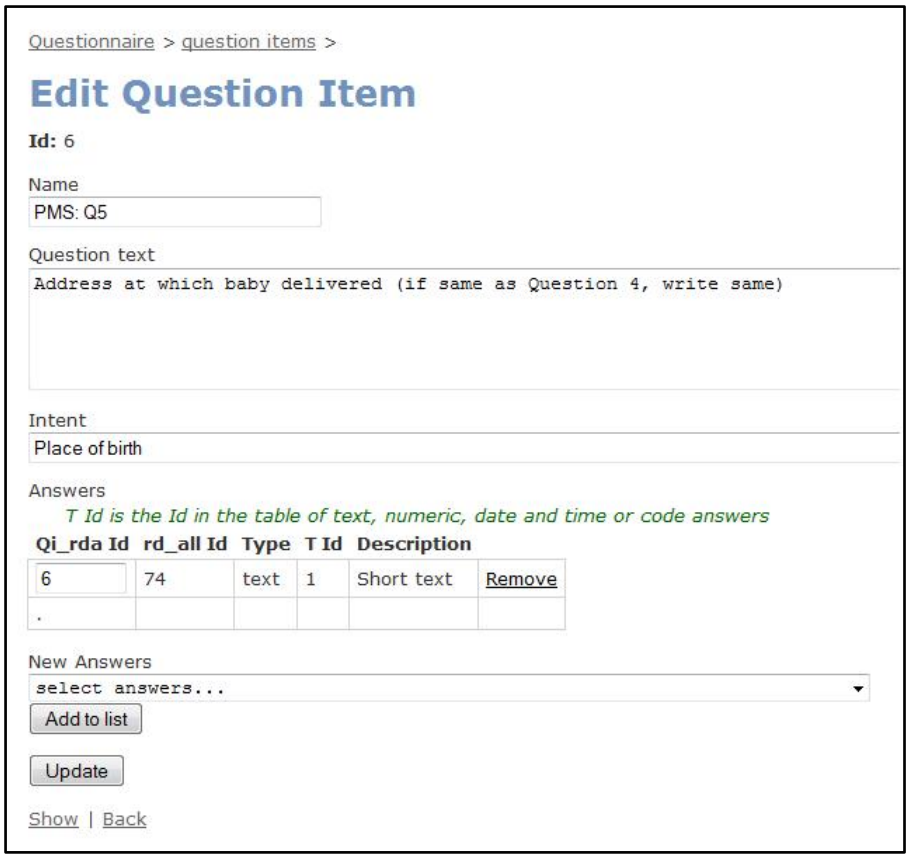


Figure 1: The CADDIES QuestionItem form

The metadata can then be exported as a DDI3 text file and imported into some other DDI3-compliant software such as Colectica for storage, linkage to metadata from survey data and further manipulation. Managing longitudinal metadata is a complex process, see Hoyle et al (2011). However, in the case of our studies, where we follow cohort members over time, there is comparatively little repetition of questions or of other DDI3 elements. This allows us to operate CADDIES on a one-sweep-at-a-time basis without excessive duplication of work.

Although the editor implements a quite limited profile designed to capture the core elements and structure of a questionnaire, later steps in our metadata management are expected to use a more extensive profile with Variables, Concepts, Harmonisation and Comparison Modules among others.

The current DDI3.1 version of CADDIES does not implement any more complex question structures than the QuestionItems and users have to split and rephrase complex questions into simpler questions. It was now felt that this might not be quite sufficient to capture the full range of question types used in paper questionnaires and that an update to 3.2 should provide users with more flexibility for dealing with tabular or complex questions. This is what this paper sets out to explore.

4 The DDI-Lifecycle 3.2 Question Structure Options

This document is based on the Public Review Version of DDI3.2. As described in *Questions – Item, Grid, and Block* (DDI Alliance) and in the online *DDI3.2 XML Schema Documentation* (DDI Alliance) there are three question structures available in DDI3.2: QuestionItem, QuestionGrid, QuestionBlock. They are all maintained within a QuestionScheme and they are referenced by QuestionConstructs *and Variables*. In the following, we limit ourselves to describing the features that are relevant to our analysis.

4.1 Question Item

As used by our editor and its DDI3 profile, the QuestionItem structure is largely the same as in DDI3.1. We use it for simple questions with a QuestionItemName for the name of the question from the questionnaire, a non-repeatable QuestionText for the text of the question, a QuestionIntent for its intent and a ResponseDomain or StructuredMixedResponseDomain if there is more than one type of response domains.

4.2 Question Block

The QuestionBlock is used exclusively in association with some evaluation material such as an image where the respondent is asked a set of questions related to the image. The frequency of occurrence of this pattern in our legacy material does not justify implementing this structure at the moment.

4.3 Question Grid

The QuestionGrid structure is used to capture tabular questions. It offers, among others, a QuestionGridName, a QuestionText and a QuestionIntent which are similar to their equivalents in a QuestionItem. It then has multiple GridDimensions which correspond to the columns in the grid. A dimension may correspond to either a list of items provided by the question through a CodeDomain or to blank cells expecting input by the respondent (Roster element). There are multiple options for the response domains depending on the complexity of the grid. Finally, the CellLabel element is also available for pre-filling cells. The QuestionGrid is a complex structure trying to cater for most eventualities and our problem is primarily a case of defining which sub-set of the structure is sufficient for our needs and realistically implementable.

4.4 Question Reference

The *SourceQuestion* in the Variable element is now replaced by the *QuestionReference*. A question reference may refer to a Question Grid. The reference needs to identify which cell(s) corresponds to the source of the variable.

5 Requirements and Issues

In order to assess the benefits of our implementation options, we need first to spell out the characteristics and features deemed desirable for our metadata management and the

pitfalls to avoid, whether these occur early on during the capture of the questionnaire metadata or later on as repercussions of the choices made for the capture phase.

5.1 An Accurate Representation of the Questionnaire

Although purely cosmetic aspects may be disregarded, we need to capture enough elements from the structure and contents of the questionnaire to be able to render an equivalent electronic version.

5.2 Meaningful Question Libraries

In DDI QuestionSchemes are seen as reusable, exchangeable and comparable components. The question structures don't have any context other than the QuestionScheme they are part of. We need therefore to create question structures at the right level of granularity. Sometimes defining a table row as a QuestionItem may be justified, in other cases a Code in a QuestionGrid may be a better option. Although not an option in our current editor, in principle it may be possible to split the QuestionScheme into multiple nested QuestionSchemes. DDI3.2 also introduces the concept of QuestionGroup which may provide an additional way of linking questions.

5.3 Meaningful Variable Sources

In DDI3.2 a variable may have a QuestionItem or QuestionGrid as a source question defined in QuestionReference. While this is sufficient information in the case of QuestionItems, for QuestionGrids we need to specify the relevant source cell and the link from the variable to the row text (usually the Code in the first Dimension) may be more convoluted.

5.4 Searching Issues

In general terms a search needs to scan not just the member fields of an item such as name, label or question literal but also very often the fields of the referenced child nodes such as the codes of a response domain. The more we make use of complex structures, the more the searchable fields are pushed further down the search trees.

5.5 Concept Issues

Concept issues are closely related to search issues. We need to be able to attach concepts to the various elements in order to facilitate the search and improve its performance while avoiding excessive duplication of information. Concepts may be collected as free text and re-assigned via a GUI menu as part of the functionality of CADDIES. Whether and how far we propagate concepts upwards from categories to questions during the metadata capture or whether we let the search propagate the concepts dynamically is still an open question.

5.6 Loops

There is the option of representing some of the tables by loops. This pushes information up towards the ControlConstructs rather than keeping it in the QuestionScheme.

5.7 Database Translators

We will need to translate the existing relational databases upon which our DDI3.1 web editor is built to a DDI3.2 version. In the first instance, we will need to be able to translate the format without necessarily making use of the new DDI3.2 functionality.

5.8 Mapping between Variables and Source Questions

The mapping between variables and source questions is entered into the repository via a many-to-many two-column text file. A variable usually has one associated question, but in the case of derived variables, multiple questions may contribute to their definitions. The mapping files may now require additional field(s) in order to specify the source cell in the case of question grids.

5.9 Harmonisation Mappings

Ideally similar tabular questions from different waves of a survey should result in a similar rendering into DDI3.2. If this is not the case, we need to allow for complex harmonisation mappings where we have a QuestionItem on one hand and elements of a QuestionGrid on the other.

5.10 Data Input Guidelines

Due to the high volume of legacy metadata to be entered into the repository, it is foreseen that clerical staff rather than DDI experts may be employed for this task. As a consequence, as far as possible, we need to define unambiguous criteria for selecting the most appropriate rendering of complex questions in order to minimise problems such as harmonisation discrepancies later on. Furthermore, the DDI3.2 profile, as implemented in our editor, also needs to avoid such a degree of complexity that the User Interface would become unusable by staff at that level of expertise.

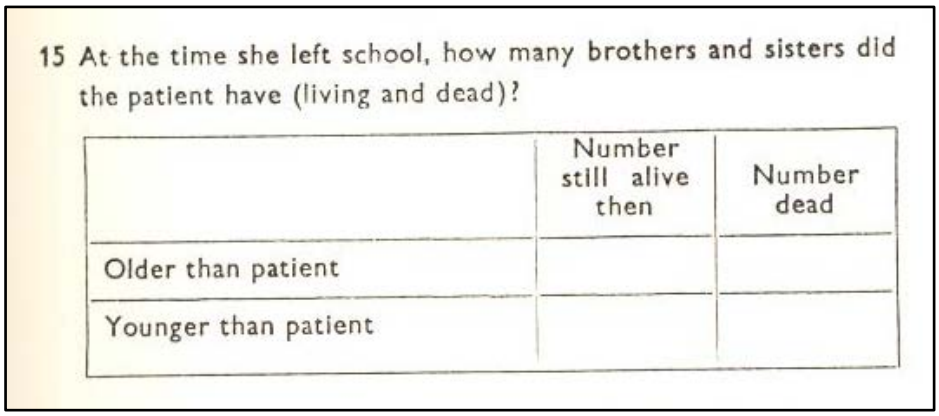
6 Examples and Discussion

In the following, we describe a couple of non-trivial sample questions from the Perinatal Mortality Survey (Butler and Bonham, 1963) in order to access the suitability of the DDI3.2 question structures to the type of material we intend to ingest if we were to implement the complete structures.

6.1 A Simple Tabular Question

Question 15 provides an example of the easier tabular questions with no real difficulties.

In this first example, the first column has some text which can be defined as a CodeList, cell[1,1] is blank, so the CodeList has to be defined without name. Column 2 and 3 have a Label in row 1 and the remaining cells have a numeric ResponseDomain. This tabular question is implementable in DDI3.2 although the diagonal symmetry, the fact that an equivalent table could be obtained by flipping rows and columns, would probably be lost.



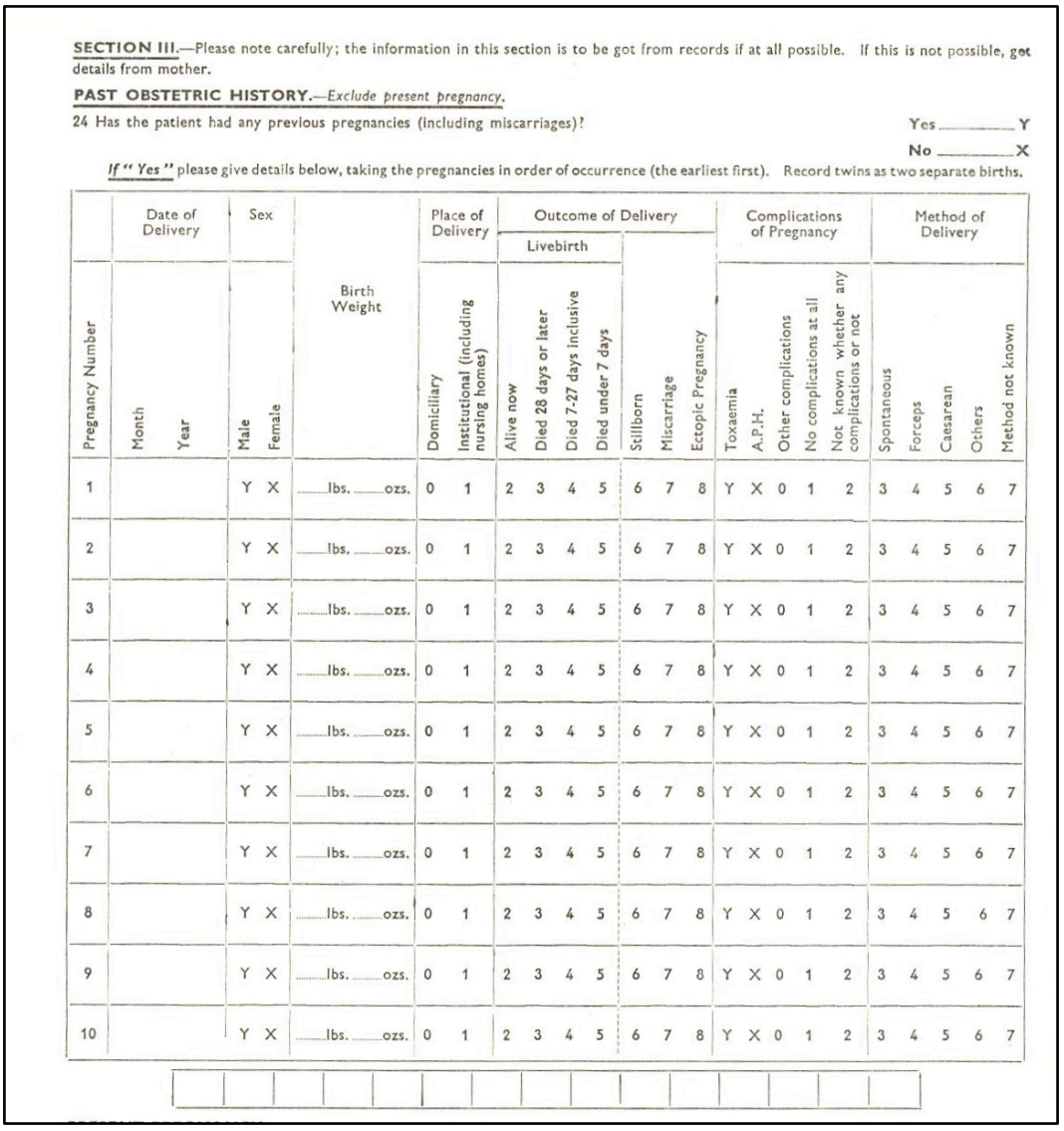
15 At the time she left school, how many brothers and sisters did the patient have (living and dead)?

	Number still alive then	Number dead
Older than patient		
Younger than patient		

Figure 2: Question 15 (Butler and Bonham, 1963)

6.2 A More Complex Table with Repeat Rows

Question 24 is more complex with a lot of information that the metadata capture needs to map to DDI3.2 elements.



SECTION III.—Please note carefully; the information in this section is to be got from records if at all possible. If this is not possible, get details from mother.

PAST OBSTETRIC HISTORY.—*Exclude present pregnancy.*

24 Has the patient had any previous pregnancies (including miscarriages)? Yes ..... Y / No X

If "Yes" please give details below, taking the pregnancies in order of occurrence (the earliest first). Record twins as two separate births.

	Date of Delivery		Sex			Place of Delivery		Outcome of Delivery							Complications of Pregnancy					Method of Delivery				
								Livebirth																
Pregnancy Number	Month	Year	Male	Female	Birth Weight	Domiciliary	Institutional (including nursing homes)	Alive now	Died 28 days or later	Died 7-27 days inclusive	Died under 7 days	Stillborn	Miscarriage	Ectopic Pregnancy	Toxaemia	A.P.H.	Other complications	No complications at all	Not known whether any complications or not	Spontaneous	Forceps	Caesarean	Others	Method not known
1			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
2			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
3			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
4			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
5			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
6			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
7			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
8			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
9			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7
10			Y	X	....lbs. ....ozs.	0	1	2	3	4	5	6	7	8	Y	X	0	1	2	3	4	5	6	7

Figure 3: Question 24 (Butler and Bonham, 1963)

In this next example, we need to ignore the purely cosmetic elements of the table and define the actual columns. The first column is a basic CodeDomain with a pregnancy index. The *Date of Delivery* and *Birth Weight* columns are in effect two columns each. *Sex* is a single column with a code ResponseDomain. *Outcome of delivery* is a single code ResponseDomain with an additional label (*Livebirth*). It is probably not critical to be able to pre-fill the response cells with the code options as in the paper questionnaire. Because the first column of the table is simply an index and the questions are otherwise uniformly repeated, this table may be better represented by a Loop from 1 to 10 with the rest of the

columns represented by a set of QuestionItems inside the Loop. There would then be a simpler DDI3 representation and an easier mapping between the QuestionItems and the Variables, but the cohesion of the question set would be removed from the QuestionScheme and shifted to the ControlConstructScheme.

6.3 An optimal implementation for question structures

The DDI3.2 schema seems likely to be able to capture most of the elements of legacy questionnaires. The difficulties reside more in the fact that a complete implementation would be excessively complex within the current framework of the software used for our editor, that the usability of the editor would require a higher level of skill than planned for such high-volume task and that the upstream management of the metadata with the linkage of question components and variables, harmonisation, comparison, search facility and related tasks may become intractable. QuestionItems on their own are probably not enough and it may be a case of implementing a QuestionGrid with limited features. For example, we may restrict ourselves to using a CodeDomain for the first GridDimension and Rosters with ResponseDomains for the following GridDimensions. It is hoped that further work on the management of the metadata beyond the capture phase may inform the requirements of the editor in more detailed way.

7 Conclusion and Future Work

We described the work of the Centre for Longitudinal Studies and the CLOSER project as a background to this analysis. We listed the surveys we intend to ingest and manage and gave a brief overview of our editor. At this stage in our work, we need to weigh up our options for the management of complex questions. QuestionGrids offer attractive features but their unconstrained use may also have negative repercussions such as an over-complex user interface for our editor and excessive requirements for the ulterior metadata management. It seems likely that we will implement a cut-down profile of QuestionGrids but the details of our implementation, the implications of our choices and the likely consequences for the survey repository are still very much work in progress.

8 Acknowledgements

CLOSER is funded by the Economic and Social Research Council (ESRC) http://www.esrc.ac.uk/ and the Medical Research Council (MRC) http://www.mrc.ac.uk/. It has been awarded a core grant of approximately £5 million for 2012 to 2017. This funding was made possible by a landmark contribution from the Government's Large Facilities Capital Fund.

9 References

Data Documentation Initiative, Questions – Item, Grid, and Block http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/drafts/Questions_DRAFT.pdf [10 Sep 2013].

Data Documentation Initiative, XML Schema Documentation http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/ [10 Sep 2013].

Amin, A., Barkow, I., Kramer, S., Schiller, D. and Williams, J., Representing and Utilizing DDI in Relational Databases, 1 December 2011, http://dx.doi.org/10.3886/DDIOtherTopics02

Hoyle, L., Castillo, F., Clark, B., Kashyap, N., Perpich, D., Wackerow, J. and Wenzig, K., Metadata for the Longitudinal Data Life Cycle, 31 March 2011, http://dx.doi.org/10.3886/DDILongitudinal03

Iverson, J. and Smith, D. (2012), Colectica: Data Management with DDI 3, EDDI12 – 4th Annual European DDI User Conference

Butler, NR & Bonham, DG. (1963) Perinatal Mortality, The First report of the British Perinatal Mortality Survey. London, E&S Livingstone.