



# **A Feasibility Study to Investigate Integrated Survey Data Collection, Fieldwork Management and Survey Data Processing Systems for Longitudinal Studies**

**Final Report**

**November 2009**

Prepared by

**Randy Banks (ULSC<sup>1</sup>), Lisa Calderwood (CLS<sup>2</sup>), Peter Lynn (ULSC), Jane Elliott  
(CLS), Geoff Angel (ULSC), and Jon Johnson (CLS)**

With the assistance of

**John McDonald (CLS)**

---

<sup>1</sup> United Kingdom Longitudinal Studies Centre (<http://www.iser.essex.ac.uk/ulsc>).

<sup>2</sup> Centre for Longitudinal Studies (<http://www.cls.ioe.ac.uk/>)

## **Acknowledgements**

The study team would like to thank everyone who has given so generously of their time to assist us, by responding to our request for comments, completing our questionnaire and, in particular, re-arranging their already busy schedules in order to accommodate the face to face meetings noted in Appendix D. In this regard we would particularly like to thank members of CHRR and the NHANES study team on whom we imposed; special thanks are due to Randy Olsen (CHRR) and Lew Berman (NHANES) for their organisational efforts.

Thanks are also due to Iram Awan, SRN Administrator at NatCen for administrative support, to Joe Lawrence and Chris Baker for their help with the diagrams in the report, and Richard Halsall and Keith Mason at the University of Essex for their expertise.

We are also grateful to the ESRC for supporting this study.

All substantive errors and omissions are the responsibility of the study team alone.

## Contents

Chapter 1: Introduction and Executive Summary .....	1
Chapter 2: Quality and Efficiency in Longitudinal Surveys.....	3
Chapter 3: The Scope for Change in the Data Production Process.....	5
Chapter 4: The Scope for Change in the Technical Infrastructure .....	18
Chapter 5: The Scope for Change in the Relationship with the Fieldwork Agency and Increasing Competition .....	21
Chapter 6. Conclusions. Towards a Common Infrastructure .....	25
References .....	29
Appendix A. Terms of Reference .....	31
Appendix B. Recommendations .....	35
Appendix C. List of Acronyms .....	37
Appendix D. Details of Meetings Held .....	38
Appendix E. Longitudinal surveys around the world .....	39
Appendix F. International Survey of Major Longitudinal Studies.....	50

# Chapter 1: Introduction and Executive Summary

## **Background**

This is the Final Report of a study commissioned by the Economic and Social Research Council (ESRC) as part of the Survey Resources Network (SRN), an ESRC investment that started in November 2008 (see <http://www.surveynet.ac.uk>).

This study aims to identify potential efficiency gains and quality improvements in the processes relating to longitudinal survey data collection, sample maintenance, data management, and the preparation and dissemination of data and metadata. Context is provided by two major ESRC investments responsible for the design, implementation, dissemination and promotion of longitudinal survey data sets that represent major resources for UK social science. These investments are:

- The Centre for Longitudinal Studies<sup>3</sup> (CLS) at the Institute of Education, University of London. Responsible for the National Child Development Study (NCDS, 1958 Birth Cohort), British Cohort Study (BCS, 1970 Birth Cohort) and the Millennium Cohort Study (MCS, 2000-01 Birth Cohort).
- The UK Longitudinal Studies Centre<sup>4</sup> (ULSC) at the Institute for Social and Economic Research (ISER), University of Essex. Responsible for the British Household Panel Survey (BHPS) and Understanding Society: the UK Household Longitudinal Study (USoc).

Both Centres underwent a mid-term review in 2007, carried out by the same reviewer. The stimulus for this feasibility study came directly from the two resulting reports, and many of the issues addressed have potentially broader implications for UK social surveys.

## **Objectives**

This study's objectives are provided by the Terms of Reference (TOR), which are reproduced in full in Appendix A. The main aims are to:

- Examine potential efficiencies in data management processes, particularly in relation to data management software;
- Examine the use of cutting-edge data collection methods for longitudinal surveys carried out at CLS/ULSC (TOR, 2.6<sup>5</sup>).

The main issues addressed here are:

- The scope of the work undertaken by CLS/ULSC, i.e. balance between the work of the principal investigator (PI) team and the fieldwork agency (TOR, 3.1);
- The supporting technical survey infrastructure (TOR, 3.2);
- The arguments for change, and the risks associated with altering the work scope and supporting technical infrastructure (TOR, 3.3);
- What systems and processes should be put in place to facilitate proposed changes and their impact on the Centres' current configurations (TOR, 3.4).

## **The Case for Change**

In this report, we use the framework provided by the TOR to consider possible changes by which CLS/ULSC might achieve quality and/or efficiency gains across the survey cycle, and the impact these might have on the relationship between CLS/ULSC and their external partners.

---

<sup>3</sup> <http://www.cls.ioe.ac.uk>

<sup>4</sup> <http://www.iser.essex.ac.uk/ulsc>

<sup>5</sup> We reference the TOR using paragraph, rather than page numbering.

The TOR focus on three important activities in the survey cycle – questionnaire programming, data quality control and data dissemination – and CLS/ULSC's main partners – fieldwork agencies and the UK Data Archive (UKDA). This report examines arguments for change in these contexts, as well as other key areas with significant influence on overall quality and efficiency.

### **Methodology**

Research for this project proceeded in two phases.

The first phase involved desk research, familiarisation with the surveys managed by CLS/ULSC and investigation of the two organisations whose practices the mid-term review reports suggested CLS/ULSC should emulate. These were the Center for Human Resource Research (CHRR), at the University of Ohio, and the National Center for Health Statistics (NCHS) in Hyattsville, Maryland, with particular respect to the National Health and Nutrition Examination Survey (NHANES), with whom on-site visits were arranged. We also met with members of the National Opinion Research Center (NORC) at the University of Chicago, who work in partnership with CHRR on the National Longitudinal Surveys (NLS), and with members of Westat, the current contractor for NHANES data collection. We also took the opportunity to meet with other US organisations and consulted in the UK with the UKDA and the National Centre for Survey Research (NatCen).<sup>6</sup>

Following our initial investigations, we published an Interim Report<sup>7</sup> in June 2009. Comment on the report was solicited by email from over 70 individuals working in survey research organisations. We received 19 replies. We then invited 18 members of survey organisations comparable to CLS/ULSC to complete a short questionnaire (see Appendix F), to which 14 responded. In addition, we organised consultative meetings with members of UK fieldwork agencies, the German Socio-Economic Panel (SOEP) in Berlin and CentERdata in Tilburg.<sup>8</sup>

### **Findings**

This Final Report is based on our Interim Report, and has been amended to reflect subsequent consultation. We received positive and encouraging feedback, both about the scope of our investigations and the recommendations made on the basis of them. Respondents provided valuable criticism both of substance and form, and were extremely helpful to us in clarifying, reconsidering and modifying our arguments, conclusions and recommendations.

Key recommendations include:

- The division of responsibility for computer-assisted interview (CAI) programming between PI teams and fieldwork agencies should not be decided *a priori* (see: Recommendation 3, p9);
- Fieldwork agencies and PI teams must share the responsibility for data quality control; ESRC should require details of proposed data quality control systems be included in proposals from PI teams for expert review (see: Recommendation 8, p12; Recommendation 9, p6);
- Many relational database management systems (RDBMS) could ultimately replace CLS/ULSC's use of scientific information retrieval (SIR); and the use of XML databases and/or XML-extended RDBMS should be considered (see: Recommendation 18, p20);
- Effective capture and re-use of metadata is central to quality and efficiency gains (see: Recommendation 2, p7; Recommendation 11, p14; Recommendation 12, p15)

---

<sup>6</sup> A member of the SRN consortium, previously contracted for data collection on the cohort studies and currently responsible for USoc fieldwork. See Appendix D for further details about these meetings.

<sup>7</sup> Available at <http://survey.net.ac.uk/sms/smsinterimreport.pdf>.

<sup>8</sup> See Appendix D for further details about these meetings.

- CLS/ULSC should establish common standards for data documentation, join the DDI Alliance, and work with the UK Data Archive to achieve DDI 3.\*-compliant documentation deposit and to enable distributed data distribution (see: Recommendation 13, p16; Recommendation 14, p16; Recommendation 15, p17; Recommendation 16, p17);
- Competition between data collection agencies for cohort and panel contracts is limited, but potential exists for ESRC to increase it, to the benefit of the studies (see Recommendation 19, p25);
- ESRC should create a common infrastructural facility to be used by CLS/ULSC and, potentially, other ERSC assets. ESRC should further consult with stakeholders to use this report as a basis for a detailed implementation strategy based on cost-benefit analyses (see Recommendation 20, p27; Recommendation 21, p27).

A full list of recommendations is provided in Appendix B.

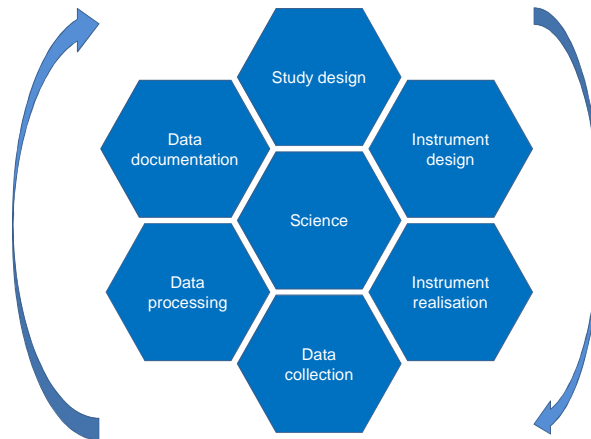
## **Chapter 2: Quality and Efficiency in Longitudinal Surveys**

This study aims to identify potential improvements to efficiency and quality in the practices relating to data collection, management and processing on longitudinal surveys. To do this, we must define the components of quality and efficiency in a longitudinal survey.

Calderwood (2009), discusses the relationship between survey processes and quality in longitudinal studies, and provides an analytical framework which we use here to evaluate the quality of existing surveys and consider how changes to processes on UK longitudinal surveys may lead to improvements in quality. Calderwood draws on a quality framework from previous work (Lynn, 2001; Statistics Canada, 1998), which maintains the main elements of quality are relevance, accuracy, timeliness, accessibility, interpretability and coherence. We argue that each wave or sweep of data collection can be conceptualised as a data production line (DPL), the main elements of which are:

- Scientific direction: deciding what data is required;
- Study design: deciding how, when and by what method these data should be collected;
- Instrument design: deciding on the detailed content of data collection instrument(s);
- Instrument realisation: producing the data collection instrument;
- Data collection;
- Data processing;
- Data documentation.

**Diagram 1<sup>9</sup>: Elements of the data production line**



This report uses the above framework to consider the scope for change in the data production processes to improve the quality of UK studies.

Improvements to quality need time and money to be realised. It is possible to improve quality in any project, but this can only be achieved either by extending existing timescales or allocating additional resources within existing timescales. Given that resources for longitudinal studies tend to be limited, the objective for these studies should be to maximise quality within the constraints in order to deliver value for money for the taxpayer, who is, ultimately, the funder. National and international bodies (e.g. National Audit Office, Eurostat) have developed frameworks with which to evaluate work they commission, to help ensure that they deliver value for money. While constraints must be taken into account when prioritising and implementing different quality measures, in this study we concentrate on the absolute value measures that we believe would have a positive impact on quality, and we do not attempt to provide the cost-benefit analyses necessary to realise them.

In a longitudinal study, there are two main elements of efficiency. The first relates to the DPL and the second to use of the data. In an efficient DPL, at each stage resources are allocated and duplication of work avoided. Efficient data use involves maximising the evidence produced from it. An efficient data source is one used extensively, by many different analysts, to produce much evidence, which has the potential to improve policy and practice and ultimately improve lives. Possible crude metrics for measuring the efficiency of a longitudinal survey are the number of publications produced using it, or dividing the total cost of producing the data by the total number of publications to produce a 'unit cost'. It is beyond the scope of this report to discuss this issue further; however, there is potential for further work on measuring longitudinal survey efficiency which focuses on the uses made of data as well as the resources invested in the survey.

In most longitudinal surveys, the different elements of the DPL are carried out by separate organisations. One of the aims here is to consider the relationship between the scientific team and fieldwork sub-contractors and, in particular, how responsibility for different elements of the survey process is allocated between them. As demonstrated in Appendix E, different studies have different ways of splitting up the DPL and the way in which this is done may have different implications for quality and efficiency. The particular arrangements that are in place are usually the result of institutional contexts, funding regulations and historical circumstances, rather than individualised consideration of the optimal arrangements for each organisation. Many different models are used on longitudinal studies around the world, from which information has been collected for this report (Appendix E).

---

<sup>9</sup> Calderwood, 2009.

Compartmentalisation of the DPL by different organisations will inevitably lead to some inefficiency of process, even if only for process-orientated reasons such as the need to put in place and monitor sub-contracts. It may also lead to some duplication of work at different points in the DPL. Nevertheless, there are potential gains from a model in which each organisation involved specialises in a different part of the DPL. This allows for investment in a specific set of skills and capabilities within an organisation, and can mean improvement in quality if each organisation applies its own quality control procedures.

The key consideration in splitting the DPL is the extent to which it will affect the quality of the end-product, i.e. the data. Other considerations in splitting the DPL include whether the experience and expertise of the organisations are used most effectively; whether their respective incentive structures sustain the quality of the data in the long-term; and whether each has sufficient control over quality control elements of its DPL section.

### **Chapter 3: The Scope for Change in the Data Production Process**

This chapter considers the different elements of the data production line and the dissemination of data and recommends changes in practice in the cohort and panel studies in order to improve quality and efficiency.

#### ***Scientific direction***

As the organisations contracted by the ESRC to create the scientific programme and deliver the supporting survey data, it is a primary responsibility of CLS/ULSC to initiate and direct the definition of the scientific scope of the surveys. However, as ESRC funds these studies to provide resources for the wider scientific community, both organisations encourage active stakeholder engagement to help determine the scientific objectives of their surveys. CLS/ULSC engage in extensive 'consultation, expert assessment and peer review' (Lynn, 2001, 3) with a wide range of potential stakeholders, including ESRC, data users, research consumers, co-funding bodies, etc. During this process of consultation, the Centres' web sites are primarily used as a means of passively publishing information on which comment or discussion is required, with the actual conversation and debate occurring in large and small face-to-face meetings, bilateral emails and telephone conversations, etc.

We think that the process by which CLS/ULSC consult the wider community could be made more effective by the use of interactive, web-enabled discussion fora. These could provide a virtual supplement to the invaluable face-to-face meetings, which are relatively expensive to organise and difficult for many interested participants to attend. CLS has already made limited, but positive use of this type of tool. In 2008 CLS launched a discussion forum on its website as a means through which to consult the wider research community on the future timing and design of the 1958, 1970 and Millennium cohort studies. This was a moderate success in that a series of 14 questions were posted on the discussion board by CLS and these received a total of 87 responses. Given the high level of engagement demonstrated by the research community, we have no reason to think that contemporary technologies that support countless lively discussion groups on the web will not support them here as well.

In addition, where more detailed comments and logging of changes are required, such as in questionnaire design, issue tracking or, more generally, collaborative development applications could provide a means by which these can be raised and addressed. Such applications minimise redundancy, as issues only need to be raised once, and maximise accountability, as the specific actions taken by the design teams to address the issues can be made public. ULSC, for example, uses Redmine to manage a number of collaborative processes within the Centre and between the Centre and NatCen, including instrument realisation. This has brought more organisation and control to all processes. We think these

benefits can and will scale up to the wider consultation processes involved in questionnaire design.<sup>10</sup>

More generally, we consider that appropriate applications could be acquired or built to support an active and widespread process of stakeholder engagement complementing existing mechanisms such as face-to-face meetings, bilateral communications, specifically commissioned expert consultations, etc.

*Recommendation 1. CLS/ULSC should investigate the use of web-based discussion fora and issue management software, as well as, more generally, of appropriate 'Web 2.0' technologies, in the process of consultation with the wider scientific community.*

### **Instrument design**

CLS/ULSC follow widespread practice and specify questionnaire instruments using word processing software, typically Microsoft Word™. Although CLS/ULSC strive for consistency when creating specifications, there are no formal guidelines either followed or shared. This can lead to ambiguities, in particular with respect to complex routing or question structures. This approach also does not facilitate the capture of structured metadata that can then be used at other stages of the data production process. For example, once questionnaire designers have input questionnaire text, variable and value labels, software could use the information to generate the script itself, produce a human-readable version of the script, create an output data dictionary, etc.<sup>11</sup>

By contrast, the Survey Research Center (SRC) at the University of Michigan has prepared extensive guidelines for producing Blaise™ specifications using Word™ (SRC, 2008, Section 2) and Westat has designed a Word™ template for specifying NHANES questionnaires that also results in a document that can be translated by software into a draft Blaise™ instrument. There are limits, however, to the extent to and efficiency with which questionnaire metadata can be captured and repurposed using word processing software. Statistics Canada is thus developing a dedicated Questionnaire Development Tool (QDT) for specification and deployment of CAI surveys, while the Demographic Surveys Division of the US Census Bureau has produced SPIDER, a web-based tool for the specification of Blaise™ instruments. (DSD, 2006)<sup>12</sup> CHRR's *Survey Suite*<sup>13</sup> design interface captures metadata that drives the process as a whole, from data collection through dissemination.

CLS/ULSC should also be using questionnaire specification tools which capture metadata in a way that is both minimally ambiguous and maximally reusable by downstream processes. CLS/ULSC cannot, however, assume that their surveys will be conducted by the same agency using the same CAI system.<sup>14</sup> The metadata models on

---

<sup>10</sup> Information about Redmine can be found at <http://www.redmine.org/>. ULSC adopted Redmine, *inter alia* as it is written using the Ruby on Rails framework, (<http://rubyonrails.org/>) which underpins ISER/ULSC's generic web architecture. Bugzilla (<http://www.bugzilla.org/>), written in perl (<http://www.perl.org/>), is another powerful and popular tracking application. For others, we recommend searching the open source Sourceforge repository (<http://sourceforge.net>) as a good start.

<sup>11</sup> A common problem when creating questionnaire specifications is maintaining the accuracy and currency of the specifications; as a matter of perceived practical expediency as deadlines draw closer, changes are often made to the instruments independently of the specifications. The Health and Retirement Survey (HRS) at the University of Michigan, which is rigorous in maintaining the currency of questionnaire specifications, is a model which would be more widely emulated if the payoffs were perceived to be greater: the greater the extent to which metadata captured during the specification process can be reused in downstream processes, the greater the motivation to treat the specification data as an authoritative description which must always be kept synchronised with the instrument(s). See also Costigan and Elder (2003) for a more general discussion about questionnaire specification and the limited extent to which the challenges involved have been addressed.

<sup>12</sup> Our thanks to Christopher J. Laskey for drawing our attention to SPIDER. (SPecialized Instrument DEvelopment Resource)

<sup>13</sup> See <http://www.chrr.ohio-state.edu/software.html>.

<sup>14</sup> Blaise (<http://www.blaise.com/>) is a popular, but not universal international CAI choice for complex government and academic surveys, but is supported in the UK only by ONS and NatCen. Major commercial agencies use a wider variety of products, sometimes in combination, including Confirmit (<http://www.confirmit.com>), Quancept

which they are based should therefore be CAI-software neutral. The DDI (Data Documentation Initiative) model will, possibly, be the most obvious choice ultimately to adopt. DDI “is an international effort to establish a standard for technical documentation describing social science data.”<sup>15</sup> The DDI standard is both open and exhaustive, the most recent version (3.\*) covering the whole of the survey cycle, thus providing an integrating infrastructure that will specifically support data documentation and dissemination (see below, p16ff). However, the standard is extremely general and, for that reason, equally complex, and DDI-based tools are still in their infancy.<sup>16</sup> Equally, as its name suggests, the DDI metadata model is built around *ex post* documentation needs, rather than the dynamic requirements of a design environment.<sup>17</sup> Alternative metadata models should thus be examined, though they should align with DDI to enable leveraging of emergent DDI-based tools.<sup>18</sup>

*Recommendation 2. CLS/ULSC should adopt and define where required common questionnaire specification standards that are independent of any particular CAI system. These standards should include the metadata models that describe questionnaires and software tools to repurpose and use the information collected during the specification process. The standards adopted should align with the open-standard DDI metadata model in order to leverage emergent tools and facilitate data documentation, dissemination and archiving.*

### **Instrument realisation**

#### *Questionnaire Programming*<sup>19</sup>

The CLS/ULSC mid-term reviews specifically suggest that CLS/ULSC should bring questionnaire programming in-house rather than contracting it to the fieldwork agencies responsible for data collection. ULSC/CLS “should prepare the instrument for delivery to the data collection subcontractor. This should be a “turn-key” application the subcontractor can load on their equipment and is ready to collect data.” (ULSC Report, recommendation 2a; see also CLS Report, recommendation 2(ii)(2)).

In-house CAI programming could improve the quality of the instruments by improving communications between questionnaire designers and programmers. Although robust questionnaire specification standards are, as already suggested, an important part of reducing programming errors, they cannot eliminate the possibility of errors completely. Specification standards provide the designer and programmer with a common language to which the former translates and articulates the intended content, and from which the latter

---

(<http://www.spss.com/software/data-collection/quantime/quancept-cati.htm>), In2Quest (<http://www.spss.com/software/data-collection/in2quest/>), SPSS Dimensions (now rebranded as part of the PASW (Predictive Analytics Software (PASW) portfolio - <http://www.spss.com/software/data-collection/> and <http://www.spss.com/uk/software/product-name-guide/>), as well as in-house systems.

<sup>15</sup> (<http://www.icpsr.umich.edu/DDI/>).

<sup>16</sup> See the *DDI Tools* site (<http://tools.ddialliance.org/>) for more information about developments.

<sup>17</sup> Note, however, Colectica, which is currently being developed as a DDI-based tool to design and document questionnaires, as well as to realise them in different CAI languages. (<http://download.algenta.com/colectica/Colectica%20Feature%20Overview.ppt>).

<sup>18</sup> QEDML (Questionnaire Exchange and Deployment Markup Language), for example, is an “is an open standard for encoding questionnaire designs” in a generic form that can be translated into the executable code used by specific CAI software, as well as “a software technology that enables the creation and deployment of questionnaires ... [w]ith QEDML, the vision of “design once” (using re-usable high level questionnaire components), and “deploy anywhere” (paper, CATI, Web, CAPI) questionnaires can become a reality.” (<http://www.qedml.com.au/>; see also Philology, 2004) Unfortunately, the QEDML metadata model remains at the time of writing undocumented and thus effectively a proprietary product. Alternatively, given the complexity and generality of the DDI model, ULSC begun in-house development of much simpler CAI-independent questionnaire specification metadata model and related tools as part of USoc developments. While different from, the ULSC metadata model is intended to align with DDI.

<sup>19</sup> See also the Responsibility for CAI programming section (below, p22), where the potential impact of in-house CAI programming on competition between data collection agencies is discussed.

interprets and translates it into an appropriate programming language. Errors may be introduced during either process of translation, and the less the designer and programmer understand of each other's professional languages, the more difficult it is to spot them. In-house programming could put the questionnaire design and programming teams into closer and more frequent contact. The latter could thus be involved in the design process, thereby gaining a better understanding of the purposes of the survey and producing an instrument that more closely represents the intentions of the designer.

Instrument development could also proceed more efficiently, as the programming resources could be guaranteed over the long-term, which is not always the case in a fieldwork agency responsible for many different projects. More generally, the design and implementation cycle could proceed more iteratively than is allowed for when the programming team is based at another institution which assigns and manages work flow. This would allow both implementation errors to be corrected, and design changes to be specified, as early as possible in the process, while the impact of either type of change is minimal. In-house programming could better ensure that the software is written to meet the data output, as well as collection requirements, as the CAI programmer would also be in closer contact with (and perhaps be identical to) the programmers responsible for post-field data processing. Finally, there would be no question of intellectual property rights in the instrument, which would make it easier to move between contractors as required.

However, these benefits to in-house programming could also, in principle, be achieved with appropriate organisational and contractual relationships between CLS/ULSC and their fieldwork contractors. For example, programmers employed by the fieldwork agency could be seconded to CLS/ULSC or, at least, be required to work on-site for regular periods and the survey designers from CLS/ULSC could be given more direct access to the programmers at the fieldwork contractor. The benefits of on-site working could be supplemented by the use of appropriate communications technology, e.g. video conferencing, virtual whiteboards, etc. There are also several important benefits of using programming staff from the fieldwork agency. The programmers employed by a fieldwork agency will work more continually with more professional colleagues on a wider variety of projects that CLS/ULSC alone could offer, thus acquiring more experience and expertise that should translate into higher quality products. Once again, however, appropriate arrangements could be made for CLS/ULSC programmers to acquire similar benefits.

Improving the quality and efficiency of the programming process does not, in other words, depend so much on the division of labour between CLS/ULSC and their fieldwork agencies as it does on 'shortening the distance' between designer and programmer, an objective that can be achieved by many different organisational means.<sup>20</sup> The pros and cons of CLS/ULSC bringing programming in-house cannot therefore be determined independently of the division of labour established with their fieldwork agencies.<sup>21</sup>

---

<sup>20</sup> We are indebted to Kymn Kochanek of NORC for this observation.

<sup>21</sup> While we accept that in an ideal world, decisions about responsibility for CAI programming would be taken in advance of the start of a fieldwork contract, this is not possible in competitive conditions where the choice of fieldwork agency, and hence CAI software, cannot be specified in advance. Nevertheless, the contraction of the CAI software market may affect the situation. Statistics Netherland continues to develop Blaise and maintain a strong international market share. However, Pulse Train Technology and Confirmit have now merged. (<http://www.confirmit.com/company/news-and-events/Press-Releases/20070917-confirmit-product-vision.aspx>) SPSS (now an IBM company) long ago added Surveycraft, In2Quest and Quantime suites to its portfolio and has been strongly marketing the Dimensions platform (now rebadged as part of the PASW (Predictive Analytics Software (PASW) portfolio (<http://www.spss.com/software/data-collection/> and <http://www.spss.com/uk/software/product-name-guide/>)) as a strategic replacement for them all. Many UK agencies have adopted or are strongly considering moving to PASW. Thus, as the range of potential CAI platforms decreases, it may become more cost-effective for CLS/ULSC to invest in acquiring the expertise in one or another as the probability that a change of agency will coincide with a change of CAI software will decrease (see also footnote 14, p6).

*Recommendation 3. Responsibility for programming the CAI instruments should be allocated after CLS/ULSC determine with their fieldwork contractors how best to achieve the highest quality instruments through the most efficient development processes.*

#### *Questionnaire testing*

Once programmed, CAI instruments require extensive testing. Complex CAI instruments pose specific challenges in this respect given the large number of possible interview vectors that the questionnaire routing can generate, and the need to validate the conditions under which they are generated. Cohort and panel instruments are intensively tested by staff at both CLS/ULSC and the fieldwork agency. As with most other studies from which we have information, the primary method used for testing involves source code review and conducting pseudo-interviews using pre-defined scenarios, a process which is both time consuming and limited in the extent to which it can trap errors based on atypical response combinations. Other, less frequently employed methods include 'data flooding' i.e. testing the instrument against randomly generated case data; HRS has developed a tool which uses the key-stroke file generated in Blaise and which allows the tester to clearly convey to the programmer the specific conditions under which the problem has occurred.

Regardless of the testing methodologies used, it is equally important that the development schedule provide sufficient time for them to be applied. Questionnaire specifications must be finalised and agreed sufficiently in advance to enable release candidates to be iteratively offered, evaluated and finally signed-off in advance of field deployment.

*Recommendation 4. CLS/ULSC should contribute to the development of new, and the improvement of existing, methods to test CAI instruments and should use current best practice for testing. The development schedule should provide sufficient time for iterative testing of instruments prior to deployment in the field.*

#### **Sample management**

Sample management includes the variety of tasks in tracing and communicating with sample members, as well as maintaining the currency and accuracy of information necessary for these tasks. Sample management is shared between CLS/ULSC and their fieldwork contractors. Simply stated, CLS/ULSC are responsible for sample maintenance between interview periods. At the outset of fieldwork, they provide current sample information to the data collection agency. During fieldwork, the sample information is again updated and then returned to CLS/ULSC at the end. This division of labour is nuanced in practice, but similar to that which exists in other comparable surveys.

As with many other organisations, CLS/ULSC use bespoke sample management systems (SMS) that have been developed in-house. These SMS have been designed around both the specific characteristics of the samples which they manage (e.g., the defining feature of cohort members is that they share the same or a similar birth data; membership of the household panels, by contrast, is based on co-residence) and particular operational requirements of the surveys (e.g., cohort samples are allocated to field on an aggregate basis, with fieldwork lasting for between 8 and 12 months, while USoc sample members are subdivided into 24 monthly samples and are allocated month-by-month across a 2-year fieldwork period).

Bespoke systems can have many advantages over general purposes systems, given that they have been designed to meet specific and constrained requirements, but because of this, they are expensive to maintain and less adaptable to change. There is no reason in principle, however, why the 'holy grail', as one informant called it, of a general-purpose SMS could not be developed for use by CLS/ULSC (and others). It would be based on a modular design, which effectively compartmentalises functionality; abstraction, which promotes generalisability; and open interfaces, which support extension beyond core tasks. To achieve this, the development process would have to take into account the sample management requirements of as many relevant surveys as possible to ensure that the specific

characteristics of one are not unintentionally mistaken for core design requirements of the many.<sup>22</sup>

*Recommendation 5. CLS/ULSC should investigate design options to enable the development of an extensible, general purpose sample management system (SMS) to support both the common and unique needs of cohort and panel surveys. SMS requirements should be established by analysing the SMS currently used by CLS/ULSC, as well as those used by other organisations, to ensure maximum generalisability.*

## **Data collection**

### *Fieldwork monitoring*

During the data collection phase, CLS and ULSC receive regular fieldwork progress reports from the fieldwork agency, providing not only details of completed interviews, but, *inter alia*, the number and distribution of booked and partially completed interviews, as well as those which have not yet been initiated. These reports allow CLS/ULSC survey managers to monitor both existing and expected trends, and to isolate and feed back response rate problems to the fieldwork agencies. Achieving and maintaining high response rates is key to the reduction errors of non-observation, of particular importance to longitudinal studies (Calderwood, 2009, p3) These reports are thus a critical element in quality control of the data collection process and provide a useful aggregate summary across the reporting period. However, they are limited in their scope and frequency and cannot answer questions that access to more continuous and disaggregated data would allow.

While not universal practice, some fieldwork agencies provide direct access to their fieldwork monitoring systems for many studies on a similar basis as internal staff and many have developed web-based tracking systems to facilitate this. These systems allow fieldwork to be monitored extremely closely at the level of individual interviewers. A good example is the PSID's 'WebTrak' system. Such systems also exist at some UK fieldwork agencies.

We believe that fieldwork agencies should provide CLS/ULSC with more frequent fieldwork monitoring information, detailed enough to measure performance at the level of the individual (anonymised) interviewer, in order to allow the PI teams to take a more proactive role in the process of minimising non-observational errors. This information could either be provided as a daily data feed or via (potentially modified) versions of the systems used by the agency staff themselves from which client reports are currently generated.

There would clearly need to be agreements between PI teams and agencies about the use and distribution of this information. This is particularly true of interviewer level data. Expected standards of interviewer performance should be agreed in advance with the PI team and should take account of existing agency-wide processes to monitor performance. Feedback on performance to individual interviewers should remain the responsibility of the agencies.

*Recommendation 6. Fieldwork agencies should provide CLS/ULSC with daily fieldwork monitoring information either as separate data feed or via the systems used by agency staff themselves subject to agreements about the use and distribution of this information.*

### *Interviewer training and data collection*

A common feature of all of the US studies examined is that, as well as a focus on the importance of response rates, there is also a strong focus on and investment in quality assurance of the data collection process from a scientific perspective. This is evidenced by

---

<sup>22</sup> Both CLS/ULSC have adapted their SMS over time to manage a wider range of surveys, but there are limits to which this process can be taken. ULSC, for example, has completely rebuilt its SMS for the USoc project. CentERdata is another case in point. 'Panelmanager' is CentERdata's third generation SMS, and though specifically designed for the MESS project, (<http://www.centerdata.nl/en/TopMenu/Projecten/MESS/>) the architecture of the system allows it easily to be adapted to other CentERdata projects. Neither CLS/ULSC nor CentERdata, however, suggest that their SMS could be used by projects run by other organisations. The difficulties involved in building a general-purpose SMS are mirrored in the fact that most CAI systems include an SMS module, but fieldwork agencies tend to reject them in favour of their own, bespoke SMS.

training sessions for interviewers which are longer, in-depth and focused on scientific objectives of the study. Interviewers are often accredited to work on the study i.e. they need to pass an examination to demonstrate their competence before they are able to do so. There is real-time checking of data from the first few interviews conducted by interviewers for quality assurance purposes, the results of which are fed back to interviewers in order for them to improve in the future. Overall, the field force is smaller, more highly trained and specialised than in the case of the cohort and household panel studies in the UK. The incentive structure and contractual arrangements for interviewers are also different than in the UK, with interviewers typically paid at hourly rates on a contract basis.

Interviewers working on the cohort and panel studies require and currently receive specialised training above and beyond the high baseline levels which are provided as standard by fieldwork agencies. Although survey-specific accreditation is not commonly done in the UK setting, this could be introduced on the cohort and panel surveys if this was one of requirements of the PI team. However, UK agencies have internal quality control procedures to ensure that interviewers understand and meet the requirements of the surveys they are working on and there is little evidence that accreditation of interviewers is needed to improve quality.

In the UK most interviewers will typically work on more than one survey at a time and the cohort and panel studies do not currently have a dedicated field force. In principle, it would be possible for the PI teams to make this a tender requirement. However, in practice, this is likely to lead to a significant increase in survey costs. There is currently no evidence that the quality of the cohort and panel studies is being compromised by the lack of a dedicated field force.

The fieldwork quality control methods employed on the UK studies are usually generic to the fieldwork agency and are not adapted to the particular requirements of specific studies. However, overall the level of fieldwork quality control employed by UK data collection agencies is extremely high. There is an international quality standard for survey research - ISO20252 – which includes quality control standard for fieldwork as well as other survey processes. Most major private sector agencies in the UK are certified to this standard. The major not-for-profit agency in the UK currently has internal quality standards which are similar to this standard and is currently working towards full accreditation with ISO20252 and the external auditing that this requires. This standard does not include any specific quality control procedures for longitudinal studies. However, UK agencies are willing and able to implement additional quality control procedures for the cohort and panel studies should they be required by the PI teams.

*Recommendation 7. CLS/ULSC should review existing interviewer training, allocation, payment structures and fieldwork quality control procedures and consider whether changes could lead to cost effective improvement in survey quality.*

### **Data delivery and quality control**

A major conclusion of the mid-term review Reports was that CLS/ULSC 'should receive the data from the field on a daily basis, load it into their database[s], and engage in real-time data quality and interview validity (that is, check for falsification) checks. The first 100 or so cases should be reviewed in detail as they come in to catch serious errors before they proliferate.' (ULSC Report, recommendation 2b; CLS Report, recommendation 2(ii)(3))

We agree that data quality control should be an important part of CLS/ULSC's roles in the data production process, ultimately responsible as they are for providing expensive and high quality data resources to the wider academic community. We do not, however, agree with the suggestion that responsibility for data quality should or does rest *either* with the fieldwork agencies *or* with the Centres:

By being more active in the receipt and examination of the data, ULSC will take on larger role in quality control. As matters stand, quality control is out sourced with the field work. Outsourcing quality control is a bad idea. By being more involved with the data ULSC will

be in a better position to detect falsification. Because field organizations compensate interviewers by the completed case, the incentives to falsify are fairly strong, especially for hard to locate and hard to contact respondents for whom traditional validation methods may fail. One can design an instrument with disguised “test” questions that may help reveal falsified cases. (ULSC Report, p 9)

Neither CLS nor ULSC have abdicated their responsibilities for data quality control, though to this point, they have met them in different ways and to differing extents. Primarily as a result of different funding histories, CLS quality control systems are less developed than those in place at ULSC. Stringent quality control systems were built for BHPS data from the outset, and they are being fundamentally redeveloped to meet the much more demanding requirements of Understanding Society. Data quality control has been costed into all bids for BHPS and Understanding Society funding and for recent bids for the funding of cohort studies.

In addition, we agree that ‘outsourcing’ quality control is a ‘bad idea’, but this is not because we think that field agencies should have no role in quality control or, even less, that agency practices necessarily contribute to a lessening of data quality<sup>23</sup> Rather, we consider that quality control is not a zero-sum game but is a responsibility that can and must be shared between PI teams and the fieldwork agencies.

CLS/ULSC quality control systems should thus not be viewed as a replacement for those that fieldwork agencies put in place themselves. Rather, we consider that fieldwork agencies *should be expected* to operate their own, high standards of quality control over all aspects of the data collection process for which they are operationally responsible. Given the complexity of the data to be collected, however, it is imperative that additional quality control mechanisms be implemented at the point of data return. These would check for both transmission errors, as well as for systematic failures in upstream processes, e.g. in questionnaire specification or programming, that may not have previously been trapped. These should include item level checks to ensure that the routing and other specified constraints are met, as well as statistical-based checks, e.g. for outliers, anomalous distributions, excessive or unexpected between wave changes, etc, any of which might indicate problems with the data rather than the reality it represents. It is good practice that additional checks be implemented independently of the data collection agencies themselves.

*Recommendation 8. Responsibility for data quality should be shared between PI teams and fieldwork agencies in order to achieve the high standards required by the cohort and panel surveys.*

To ensure that data quality control systems are given proper weight in project planning and operations, we consider that they should be specifically described and be a line-item cost in all bids to ESRC for strategic data resources. To assess the proposals properly, we also consider that at the least one referee should have extensive data management expertise, and be specifically asked to comment on them.

*Recommendation 9. ESRC should require costed details of data quality control systems as part of tenders for strategic data resources to which CLS/ULSC and others might respond. At least one referee for the bids should possess the technical expertise required to evaluate the proposals.*

As the Reports suggest, CLS/ULSC do not acquire and process data or paradata (data about the data collection process) in real-time or, more practically, on a daily basis.<sup>24</sup> Nor are

---

<sup>23</sup> Our experience and knowledge of major agencies is that they are committed delivering high quality data, and therefore see the value of effective quality control procedures.

<sup>24</sup> CLS receives a number of batches of ‘test-data’ during fieldwork but a single main data delivery of coded data occurs as a single transfer after fieldwork has closed. With BHPS, ULSC received batches of data throughout the fieldwork period, the frequency of which varied according to the intensity of interviewing, but only rarely exceeding one delivery per week. USoc data delivery specifications require data delivery at the close of each month’s fieldwork, with provision made for more frequent deliveries of test data at the outset of a wave. (USoc

they alone in not receiving data on a daily basis,<sup>25</sup> though many other studies do, and UK fieldwork agencies do not see any reason in principle why this could not be achieved for the cohorts and panels. Continuous data delivery would enable CLS/ULSC to improve quality by implementing a 'responsive survey design'.<sup>26</sup> It could also help shorten the delivery schedule for releasing data to the secondary analysis community, thus contributing to the *timeliness* dimension of overall quality.<sup>27</sup>

There are nevertheless limits to the extent to which daily data deliveries and continuous quality control can be implemented: some types of data do not lend themselves to continuous data delivery, while some quality control checks must wait until the end of a fieldwork period<sup>28</sup> There are also obstacles to data release that cannot be overcome by more frequent data deliveries.<sup>29</sup> Despite these provisos, we consider that CLS/ULSC should aim for daily delivery of data.<sup>30</sup>

*Recommendation 10. Fieldwork contractors should provide CLS/ULSC with daily data and paradata delivery. CLS/ULSC should develop quality-control and data throughput systems which use this data on a continuous basis.*

The benefits of daily data delivery will be reduced to the extent that deliveries or the use which can be made of them do not commence well into the field process. In order for this not occur, CLS/ULSC and their contractors must be prepared at the outset of the field process to transfer data that can be incorporated into processing systems.

Creating a transport mechanism to enable this is a necessary, but trivial requirement.<sup>31</sup> Defining the content, structure and format of the data to be transferred in good time to enable immediate transmission and use is also required, will be less likely if current practices continue. At the moment, contractors tend to use their CAI system's default output facilities to create data files, in SPSS™ format, for the most part. The difficulty with this procedure is that even with a certain amount of structuring of the output file, into 'household' or 'family' and 'individual' level files, default output files do not always transparently relate back to the questionnaire specification, reflecting both programming and system idiosyncrasies; the more complex the questionnaire, the more this is the case. CLS/ULSC can thus spend a

---

waves have a 24 month fieldwork period, with the sample allocated and administered on a month-by-month basis)

<sup>25</sup> Apart from interim test data, the highly respected German Socio-Economic Panel also receives a single data delivery at the end of the field process.

<sup>26</sup> "The ability to monitor continually the streams of process data and survey data creates the opportunity to alter the design during the course of data collection to improve survey cost efficiency and to achieve more precise, less biased estimates. We label such surveys as 'responsive designs'" (Groves & Heeringa, 2006, 439).

<sup>27</sup> Despite the complexity of the project, real-time data delivery allows NHANES data to be released to secondary analysts within six months of the close of fieldwork.

<sup>28</sup> Coding of open-ended responses and key-to-disk of paper questionnaires, for example, are both more efficiently accomplished in batch mode. Household panels such as BHPS and USoc would not necessarily want individual data delivered as soon as it becomes available, but might (as is the case with BHPS and USoc) prefer to wait until interviews with all individuals in a household are complete. Similarly, checks that there is returned data for all sample members who were allocated to field cannot be done until the field process is nominally complete as prior to this all such checks will, by definition, fail.

<sup>29</sup> Weighting and imputation, for example, must wait for the complete data set.

<sup>30</sup> As with the use of fieldwork monitoring data, the process by which the scientific teams' quality control systems will feedback and influence the data supply chain must be agreed with the data contractors.

<sup>31</sup> For example, an application at the CLS/ULSC end could daily probe for and fetch data archives from a secure NatCen URI based on an agreed file naming convention that includes a variable date component. ULSC has already implemented a secure transport mechanism for Understanding Society via the Understanding Society web 'portal', which is used for most communication with NatCen, including data delivery. Because it was designed to handle batch deliveries, however, the portal is a 'push' mechanism, in that NatCen manually initiates the transfer. It could, however, be adapted to become a 'pull' mechanism as described above.

great deal of time comprehending and importing the data into local systems, as well as in subsequently adapting them to any wave-specific needs.<sup>32</sup>

To overcome these problems, data output specifications need to be well defined in advance of the commencement of data delivery. They should be created simultaneously with the specification of the instrument itself, and should be generated from metadata collected at this stage. Instrument programming can then directly take into account data output requirements, which can then be tested and verified simultaneously with testing data collection. Staff responsible for data management should also be directly involved in questionnaire specification to ensure that the logic of the data collection instrument is aligned with output requirements. Specifications should be produced, and based on common standards which are independent of the CAI system from which the data will be output and are neutral with respect to the target systems.

*Recommendation 11. CLS/ULSC should adopt common data output specification standards that are CAI- and data management system-independent. These standards should be used to specify the content, structure and format of all data to be delivered. Specifications should be generated from metadata collected at the questionnaire specification stage to allow data delivery requirements to inform questionnaire programming and be tested simultaneously, well in advance of the commencement of fieldwork. Data management staff should be directly involved in questionnaire specification to ensure that the logic of the data collection instrument aligns with output requirements.*

Although we are not in a position to provide precise details of what is required, we suggest that use of the XML-based triple-s survey data interchange standard be investigated.<sup>33</sup>

### **Data processing and data documentation**

As the TORs (3.1) note, “At the heart of the mid-term reviews of the Centres, is a suggestion that it would be beneficial for CLS and ULSC to have greater control over all the survey process from design through to data dissemination” and, as the mid-term reports make clear, the effective capture and use of the metadata that describes all aspects of this process is central to achieving this objective (ULSC Report, 9ff; CLS Report 21ff).

In the case of a survey, even a minimal list of relevant metadata is large and complex,<sup>34</sup> and the various items of which it is comprised, as well as the relationships between them, are used at different points in the survey process. The content of a question, the order in which it is asked and the conditions under which it is asked, for example, form part of the questionnaire specification and the questionnaire itself. This information is subsequently used by quality control processes, which check that question responses are valid given the conditions in which they were asked. The same information is included in the end-user documentation used by the secondary researchers to determine the analytical substance and significance of a particular variable.

---

<sup>32</sup> Systems such as Survey Suite used by CHRR (<http://www.chrr.ohio-state.edu/software.html>) do not suffer from these problems as they integrate questionnaire design, data collection and data processing. Given the UK situation, however, we do not consider it possible to impose a particular data collection tool on fieldwork contractors and work on the assumption that CLS/ULSC must be able to interface with the variety of systems used by potential contractors. See Responsibility for CAI programming, below p22ff.

<sup>33</sup> See <http://www.triple-s.org/> for full details about triple-s. triple-s is a very simple but effective standard for data interchange, which ULSC currently uses in the USoc project.

<sup>34</sup> For example, the mid-term reports suggest that the metadata required to describe a question in a longitudinal survey at the very least includes: “The question identifier and the title(s) associated with the variable representing the question’s answer with the facility to connect the same question asked in different sweeps ... Descriptors that characterize or index the content of the question ... The question text ... check items that [lead] into the question .. [the] allowable responses to the question and data specifications for [them] ... Routing instructions ... Real-time edit specifications ... Pre-loaded values ... Text fill specifications ... interviewer and respondent [instructions] ... Alternate language versions ... Date and time stamps ... comments about the accuracy or interpretation of the item or its source ... Notes to the support staff about complexities associated with the question Links to supporting documentation ...” (ULSC Report, 10; CLS Report, 21-22)

Capture of this metadata in machine-parseable form is thus an important aspect of gaining control over the survey process as a whole and increasing the efficiency and quality of it. Systems design based on common metadata repositories lead to effective integration of the process, as multiple applications can use, or copy without loss, a single instance of the metadata for its own purposes. This eliminates the need for time-consuming and error-prone manual transfer of the metadata between applications. It also allows generic systems to be built which can readily adapt to the variety of differences between different cycles of the same survey, as well as between different surveys, by changing the content of the metadata used rather than having to (re)develop the systems themselves.<sup>35</sup> The CHRR survey processing software, and the systems on which the NHANES delivery chain are based, represent the most effective and exhaustive use of this design methodology of which we are aware.<sup>36</sup>

We thus wholeheartedly concur with the mid-term reports about the need for CLS/ULSC to base their survey management systems around common metadata repositories that can be used by the components of which they are comprised. The metadata model on which they are based should be generic enough to represent the common elements and processes involved in any longitudinal survey, but must also incorporate the specificity to represent the differences between them, as well as sufficient extensibility to ensure that they can accommodate future requirements.<sup>37</sup> To ensure future-proofing, the metadata model should also be independent of specific technical solutions that might implement or use it. The model should also align with other data exchange and documentation standards, such as DDI and SDMX<sup>38</sup> to facilitate transport, archiving and dissemination of the output products.

We are not in a position to state the extent to which existing metadata models could be used off-the-shelf or, if not, which elements could profitably adopted. Those developed by CHRR or the NHANES project are intimately linked to a specific technology (CHRR) or survey (NHANES), and might not be generalisable to the extent required by CLS/ULSC. These and other metadata models, the DDI 3.\*, in particular, must be considered in more detail to determine the extent to which they could be used to support USLC/CLS relevant needs.

*Recommendation 12. CLS/ULSC should develop a generic metadata model on which a common repository can be developed to support survey processing systems. This model should be extensible and implementation independent, incorporate both the common and specific features of the surveys they currently manage, and align with other relevant metadata standards. The development process should investigate and, as possible and appropriate, use existing metadata models supporting comparable projects.*

---

<sup>35</sup> As the mid-term reports put it: "Every time the data or instrument changes hands, misunderstandings and errors will occur. These projects are so complex and large that every time someone passes responsibility for the instrument or data to someone else, something unexpected will occur. The best protection against this sort of human error is to keep a single integrated archival system which every step of the process references and uses." (ULSC Report, 12; CLS Report, 24)

<sup>36</sup> NHANES provides a quasi-experimental example of the benefits derived from metadata-driven systems, having reduced its data release cycle from 3.5 years post fieldwork, for the periodic pre-1999 surveys, to 6 months, for the continuous surveys that have since been running.

<sup>37</sup> The cohort studies, for example, sample and track individuals based on shared birth dates, and change over time the measures employed, the means by which the data are collected, and the relevance of other members of their household, according to the stage of the life course reached by a cohort member. The panel studies, by contrast, sample individuals based on their household membership, and track them, as well as other individuals with whom they establish new household relationships, using repeated measures as much as possible. Equally, CLS/ULSC envision a possible role for non-traditional question stimuli and data collection e.g. (pre-)recorded audio/visual data.

<sup>38</sup> On DDI (Data Documentation Initiative), see <http://www.icpsr.umich.edu/DDI/>. On SDMX (Statistical Data and Metadata Exchange), see <http://sdmx.org/>. See above, p7, for a brief discussion of the pros and cons of DDI.

<sup>40</sup> See <http://www.data-archive.ac.uk/>

The TOR (3.6) requires that “The Study should also consider the work of the PI teams in preparing data for deposit at the UK Data Archive and potential efficiencies which could be gained through preparing metadata to common standards.”

DDI 3.\* is the most recent version of the international documentation standard adopted by data archives throughout the world. Unfortunately, the standard is very complex and sufficiently new that the UKDA and others are still in the process of rolling it out to support internal processes. As “a centre of expertise in data acquisition, preservation, dissemination and promotion and ... curator of the largest collection of digital data in the social sciences and humanities in the UK ... [which] ... now houses several thousand datasets of interest to a wide range of researchers,”<sup>40</sup> the UKDA is committed to interface with a very wide variety of data depositors who provide data and documentation produced in a diverse range of formats. The UKDA is currently unable to receive data described by DDI 3 mark-up, and the scope, in consequence, for efficiency gains in this respect is extremely limited.<sup>41</sup> However, the UKDA are working towards, in the medium term, implementing the technologies to ingest data packages marked up in DDI 2 and 3.

Our recommendations concerning the development of a metadata model are intended to ensure that ULSC/CLS will be able to make use of common metadata standards for deposit when this is practicable. As the UKDA is currently engaged in initiatives to allow deposit through the use of DDI compliant inputs we recommend that CLS/ULSC work closely with the UKDA to develop and take full advantage and promote the value of these initiatives.

*Recommendation 13. CLS/ULSC should work closely with the UKDA to achieve DDI-compliant metadata deposit.*

We also consider that CLS/ULSC should have more direct input to the development of the standard through membership of the DDI Alliance<sup>42</sup>. Unlike previous versions of DDI standard, DDI 3.\* is sufficiently complex to represent the generic structural aspects of cohort and panel data. However, this includes only limited semantics for the specific markup of longitudinal surveys. Membership of the Alliance would enable CLS/ULSC to contribute both to the development of the standard to better represent cohort and panel data, and to increase their understanding of its use and benefits.

*Recommendation 14. ESRC should fund at least one of CLS/ULSC to become a member of the DDI Alliance, specifically to contribute developments of benefit to longitudinal surveys in general, and the cohort and panel surveys in particular.*

Metadata capture and exchange through compliant standards is not, however, only a necessity in the completion of the survey life-cycle, the end products of which are the data and documentation made available to the secondary analyst. Longitudinal data, in general, is intrinsically more complex than cross-sectional data as it by definition includes a third, time, dimension. The data collected by the cohort and panel, as well as other longitudinal studies, add to this complexity by introducing multiple data streams and units of analysis – households, families, event histories, etc – with linkages between them that are not always immediately transparent.<sup>43</sup> Although, as the mid-term reports suggest, “[i]t is essential that

---

<sup>41</sup> We note that the Data Archive is not alone in this respect. ICPSR, for example, although the institutional host for DDI, nevertheless considers that a DDI 3.\* deposit “would be a challenge” that ICPSR would welcome, but could not, at the moment, meet. However, CentERdata is making effective use of a subset of the DDI 3.\* object model in the ‘Questasy’ system to document LISS (Longitudinal Internet Studies for the Social Sciences) data. ([www.lissdata.nl](http://www.lissdata.nl)) and the first Annual European DDI Users Group meeting will shortly take place. (<http://www.iza.org/eddi09>)

<sup>42</sup> The DDI Alliance is responsible for the development and ratification of the standard. See <http://www.ddialliance.org/org/>.

<sup>43</sup> BHPS and Understanding Society, for example, are both based on the same underlying sample structure in which households are defined on a wave-by-wave basis based on sample members who share common residential characteristics. At any given wave after the first, in consequence, there may be implicit relationships between sample members in different households which can only be established for research purposes by navigating all previous data waves.

the data file preserve the relationships between the data and the sampling structure and respondent relationships.” (ULSC Report, 9; CLS Report, 21), the data may nevertheless be presented to the secondary analyst in a variety of ways designed to meet different analytical requirements.<sup>44</sup> Extensive documentation is thus required to clarify both the content of and relationships within the data and between different datasets. CLS and ULSC have adopted specific approaches to this problem.<sup>45</sup> Neither solution is ideal, and the lack of consistency between studies is unsatisfactory to the extent that it imposes additional burdens on the potential data user. CLS/ULSC should therefore consider best practice in the approaches adopted by other, comparable longitudinal studies, establish common standards for the distribution and documentation of longitudinal data.

*Recommendation 15. CLS/ULSC should establish common standards for distribution and documentation of longitudinal data based on best international practice.*

### **Data dissemination**

One major difference between CLS/ULSC and some other comparable studies, is that the latter provide direct access to much of their data. This enables them to provide more flexible access to the data, e.g. subsetting capabilities, and to better integrate it with the available documentation.<sup>46</sup> Historically, CLS/ULSC, have, however, relied entirely on the UKDA for both data preservation and dissemination. In certain respects, this division of labour is extremely efficient, as CLS/ULSC can rely on UKDA’s generic authentication system to ensure data access is correctly regulated. However, it limits the extent to which CLS/ULSC can enhance documentation and provide flexible data access opportunities designed to meet the specific needs of cohort and panel data users. ICPSR offers data to users on some studies in different formats as selected downloadable variables or as sweep by sweep data blocks more analogous to the UKDA, though its experience is that users rarely utilise the former. This should be borne in mind before investments are made in one area which could better serve the user community if placed elsewhere.

Advances in web-based technologies now make it possible for the UKDA to separate part of its authentication role from its dissemination function, allowing CLS/ULSC to create access systems best suited to the needs of the cohort and panel studies.<sup>47</sup> Consequently we agree with the mid-term reports that CLS/ULSC should “re-examin[e]... how the data are distributed to public users” (ULSC Report, 12; see also CLS Report, 24), and we propose that CLS/ULSC should work with the UKDA to explore possibilities for the creation of tools which are specifically targeted towards the specific needs of cohort and panel study data users in a way which does not undermine impact significantly on the preservation function provided by the UKDA. Such tools would enable CLS/ULSC in particular to create and distribute a greater variety of data products and directly integrate them with documentation down to the item level.

*Recommendation 16. CLS/ULSC should work with the UKDA to explore options for data dissemination that provide both centralised user authentication and distributed access to the data and documentation through systems specifically designed to meet the needs of data users.*

---

<sup>44</sup> For example, the data may be restructured to present the data aggregated or distributed to different levels of analysis and/or data may be subsetted to create topic-relevant datasets.

<sup>45</sup> See the CLS data dictionary (<http://www.cls.ioe.ac.uk/datadictionary/>) and BHPS documentation (<http://www.iser.essex.ac.uk/bhps>)

<sup>46</sup> ULSC did begin a project to integrate BHPS data with the existing documentation to provide data extraction, restructuring and subsetting capabilities, but the pressure of operational demands did not allow progress beyond a preliminary prototype.

<sup>47</sup> One possible model might be for the “UK Data Archive could continue to authenticate users but when it is time to push the data out to users, it could employ an encrypted link to the master data base to generate customized extracts. With the Web it does not matter where the server is located.” (ULSC Report, 12; CLS Report, 24)

In addition to creating, documenting and disseminating high quality data. CLS/ULSC must also ensure that secondary researchers are best equipped to use the available resources. Both CLS/ULSC have produced teaching data sets and organise well-attended training courses, alone and in conjunction with, *inter alia*, the UKDA, the Essex Summer School in Social Science Data Analysis. We believe, however, that CLS/ULCS can and should, like other studies, develop web-based user training resources following best international practice.<sup>48</sup>

*Recommendation 17. CLS/ULSC should continue to develop existing mechanisms for user training, and should make significant investments in creating web-based training resources following best international practice.*

## **Chapter 4: The Scope for Change in the Technical Infrastructure**

The TOR require that we ‘consider two major issues highlighted in the CLS/ULSC mid-term reviews: the scope of work undertaken at the two Centres and the supporting technical survey infrastructure. These two issues are strongly linked.’ (3.1) In this chapter we consider questions surrounding the future technical infrastructure of CLS/ULSC, in the next chapter, we consider the scope of work undertaken at the Centres and by fieldwork agencies.

### **Relational Database Management Systems**

The mid-term reviews consider that SIR™ is ‘outdated’ and should be replaced by ‘modern’ relational database management systems, (RDBMS) which “are the most effective tools for work with, and archiving, longitudinal survey data.” (TOR, 3.2) Though CLS/ULSC have long-term experience with SIR, both recognise that there are many reasons to achieve SIR™’s functionality by other means. Thus, as the ULSC mid-term review notes, “ULSC is examining its database options” and encourages “... them to consider their longer term strategy for handling the data flow before they pick a RDBMS vendor and consider how that choice will fit in with their longer term plans.” (ULSC Report, 8)

CLS/ULSC currently use SIR™ as a strategic database management system (DBMS) for post-field storage and processing of survey data, in particular, quality control and creation of data output products. SIR™ is well suited to these tasks for at least two reasons. Firstly, unlike other RDBMS’, SIR™ databases can better represent the ‘natural’ structure of survey data.<sup>49</sup> Secondly, the primary interface to SIR™ databases is provided by the proprietary programming language, PQL™ (Procedural Query Language) rather than is typically the case with other RDBMS, which use the open-standard SQL (Structured Query Language)<sup>50</sup>. PQL™, unlike SQL, is a feature-rich, but equally general purpose programming language that is particularly suited for manipulating data in the aggregate. However, if PQL™ is one of SIR™’s strengths, it is also a weakness. PQL™ provides the only effective interface to SIR™ databases. Although SIR™ has always been marketed as a general purpose RDBMS, it is particularly suitable for the management of research and scientific data,<sup>51</sup> a niche market

---

<sup>48</sup> See for example the NHANES tutorials at <http://www.cdc.gov/nchs/tutorials/>.

<sup>49</sup> For example, in building systems to manage NHANES data, “[t]he concept pursued for operational data extraction was the use of a fully relational data handling method for unloading actively needed data from Blaise™ instruments. This is not a new concept in general IT methods. Yet, it is not often applied to Blaise™ data since these data do not readily yield to a fully relational structure. Native Blaise™ structure is essentially horizontal with many fields of data on few rows. A fully relational structure is essentially vertical with few fields on many rows, thus reducing data redundancy. Therefore, reducing the columns and increasing the rows better leverages the manipulative potential of the data, offers greater flexibility of data combinations, and is more change tolerant.” (Hill et al, 2001, 13)

<sup>50</sup> The current version of the SQL standard is defined by ISO/IEC 9075-1:2008 ([http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=45498](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45498))

<sup>51</sup> SIR™ was originally developed as a complementary product to SPSS™, as well illustrated by the many similarities between PQL™ and SPSS™ ‘syntax’. SIR™’s target audience was indicated by it’s choice of name, which, in the days WAWA (when acronyms were acronyms), ‘SIR’ stood for ‘Scientific Information Retrieval’

which SIR™ can no longer dominate.<sup>52</sup> PQL™ expertise is thus increasingly difficult to acquire and, despite its many benefits, CLS/ULSC should not, in consequence, continue to rely on it as the primary data storage platform. The mid-term review reports recommend that CLS/ULSC replace SIR™ with another RDBMS on which a closely coupled suite of applications can be built to manage the survey process in its totality, from questionnaire design through to data dissemination, most directly akin to the systems that CHRR has developed. As the reports suggest, this would have positive impacts on both the quality and efficiency of the process as a whole.<sup>53</sup>

While we agree with the basic thrust of these conclusions, we believe they must be qualified in a number of significant respects. As is discussed more fully in the next chapter, we do not think it desirable for CLS/ULSC to impose CAI software on a fieldwork contractor. Nor, by contrast, do we think that it would be desirable for CLS/ULSC to employ in-house systems that are closely coupled to those used by fieldwork agencies, as this would make it exceedingly difficult to establish relationships with new agencies. Secondly, even if CLS/ULSC were to adopt “a relational data base as the core software tool for the documentation and data dissemination tasks” (ULSC Report, 12| CLS Report, 24), there will remain a greater or lesser requirement to interface with applications which do not use the database contents directly. Thirdly, we recognise that the use of an RDBMS cannot be a complete solution to the diverse functionality that CLS/ULSC requires. For instance, in the organisations that we have investigated, complex data manipulations, such as cross sweep data cleaning or generation of anything other than the most simple derived variables, are often performed outside of the RDBMS and the results imported back. This functionality, currently achieved inside SIR, will have to be achieved in other ways.

Fourthly, CLS/ULSC should consider the viability of alternatives to RDBMS technology, XML databases in particular. As Harold (2005) notes, “Relational databases in general, and SQL databases in particular, have been so incredibly successful that they've almost completely eliminated the competition, at least in mind share if not always in actual installations.” An RDBMS provides extremely flexible data management capabilities based on the mathematical simplicity of the underlying relational data model.<sup>54</sup> An RDBMS can, in consequence, represent any collection of data regardless of its specific semantic content. It is not, however, always the best solution: “When your only tool is a hammer, everything looks like a nail. When your only tool is a relational database, everything looks like a table. Reality, however, is more complicated than that. Data often isn't tabular and can benefit from a tool that more closely fits its natural structure” (Harold, 2005). XML markup and related technologies provide an increasingly widespread choice of alternative tools. An XML document can carry the same information and more as can an RDBMS.<sup>55</sup> XML documents, however, are typically stored as files or streamed for the purpose of data exchange, and manipulating large documents with standard XML tools can carry significant performance overheads compared to the manipulation of similar data in an RDBMS. XML databases, or RDBMS' with XML extensions, are being developed to address this deficit, though expert opinion is that significant development is required before XML databases achieve the cost-performance ratios of RDBMS.<sup>56</sup> Nevertheless, we consider that XML, and XML database

---

<sup>52</sup> On the one hand, statistical packages like SAS™ and SPSS™ have vastly increased their native data management capabilities and can, in a variety of ways, directly interface with RDBMS'. On the other hand, organisations have increasingly consolidated their RDBMS support in favour of the general purpose RDBMS vendors.

<sup>53</sup> “The disjointed strategy for moving instrumentation and data between ULSC and its contractor requires more resources in total to support. The same is true of the exchange between ULSC and the UK Data Archive', both of which are extremely prone to human error. (ULSC Report, 9, 12; CLS Report; 21, 24)

<sup>54</sup> The seminal exposition of the relational data model is provided by Codd, 1970.

<sup>55</sup> For example, 'order' is an important element in surveys, e.g. as concerns the sequence in which questions are asked. The relational data model is based on set theory, which has no intrinsic concept of 'order', either between tables, rows or columns. Sequencing of elements is an inherent feature of XML documents.

<sup>56</sup> See Williams (2005) for an analysis of the performance of XML compared to relational databases.

technology in particular, will provide a viable alternative to RDBMS technology in the near future, and should be considered by CLS/ULSC as part of the retooling process.<sup>57</sup>

RDBMS technology will doubtless continue to have a significant place in the storage environments of many organisations, CLS/ULSC included. The studies and organisations we have consulted thus far use a wide variety of RDBMS, including Oracle™, Sybase™ and SQL Server™, and we are not in a position to recommend any one product for use by CLS/ULSC. However, in selecting a product, performance measures<sup>58</sup> and a good support for the SQL standard<sup>59</sup> should be taken into account. In selecting an RDBMS to replace SIR™ and extend the use of RDBMS technology, CLS/ULSC may wish to reconsider the current use of other general purpose RDBMS platforms. CLS, for example, uses SQLite<sup>60</sup> for meta-data handling, and some limited data storage, and has utilised MS Server™<sup>61</sup> for its Cohort Maintenance functions for several years. BHPS sample management was originally based on MS/Access™, and has since been completely redeveloped for Understanding Society, with the backend database provided by the open source PostgreSQL™ product.<sup>62</sup> Sample maintenance applications accessed by in-house and fieldwork staff are integrated into a web-based 'portal', which also supports secure data and document exchange between ULSC and NatCen. Both data, and the metadata used to drive the application subsystems which comprise the portal, are stored in PostgreSQL™ databases. CLS/ULSC should consider the extent to which these platforms can be used for other purposes, thus leveraging existing expertise, as well as the costs of replacing and/or continuing with them should an alternative solution be adopted.

*Recommendation 18. In selecting an RDBMS to replace SIR and extend the use of RDBMS technology, CLS/ULSC should prioritise platform-independent, open-source RDBMS that provide good support for the SQL™ standard. CLS/ULSC should take into account their use of existing general purpose RDBMS platforms, and the extent to which their use is application dependent and/or can be generalised. CLS/ULSC should also investigate the extent to which RDBMS technologies can be replaced with XML and XML databases or XML-enabled RDBMS.*

### **Legacy Data<sup>63</sup>**

The TOR asks that we “consider mechanisms for incorporating legacy data from previous waves of the longitudinal studies.” (3.4) Details of how this would be accomplished cannot be suggested until the development process is further advanced. We do not foresee that these will pose particular or difficult problems in principle, though there could be considerable resource implications for incorporating legacy data into any new system as all the UK studies, and the birth cohort studies in particular, have considerable volumes of

---

<sup>57</sup> Harold (2007), for example, suggests that XML database development is at the level RDBMS' had achieved in 1994 e.g. 24 years after Codd's (1970) original formulation, while this statement was made less than 10 years after the first XML standard was ratified. (W3C, 1998)

<sup>58</sup> RDBMS performance measures typically focus on transaction throughput (Transaction Processing Performance Council (<http://www.tpc.org>), which will be of only limited use to CLS/ULSC, who will also be concerned with aggregate processing performance. See ([http://en.wikipedia.org/wiki/List\\_of\\_relational\\_database\\_management\\_systems](http://en.wikipedia.org/wiki/List_of_relational_database_management_systems)) and ([http://en.wikipedia.org/wiki/Comparison\\_of\\_relational\\_database\\_management\\_systems](http://en.wikipedia.org/wiki/Comparison_of_relational_database_management_systems)) for more general information about the major products and their characteristics.

<sup>59</sup> Good support for the SQL standard (ISO/IEC 9075-1:2008, [http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=45498](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=45498)) will ensure that CLS/ULSC is not inextricably tied to one product or another. Many database vendors have added their own proprietary extensions to SQL. However useful these might be, they should not influence the selection of an RDBMS product.

<sup>60</sup> SQLite™ is available from <http://www.sqlite.org/>.

<sup>61</sup> Information on MS SQL Server™ is available at <http://www.microsoft.com/sqlserver/2008/en/us/default.aspx>.

<sup>62</sup> PostgreSQL™ is available from <http://www.postgresql.org/>.

<sup>63</sup> I.e. data collected in previous survey waves which has been processed by systems and distributed in ways which are or may no longer be directly supported.

legacy data of different kinds and all metadata would have to be mapped on to the new database structure.

## **Chapter 5: The Scope for Change in the Relationship with the Fieldwork Agency and Increasing Competition**

The TOR specifies that we should review our existing relationships with fieldwork agencies and review the scope of work undertaken by the Centres. The TOR state that:

The mid-term reviews argue that the use of relational database management systems (RDBMS) holds the key to reforming the balance of work between the academic PI teams based at the Centres and the fieldwork agencies. ... consider the relationship between the academic PI team and fieldwork agency, if RDBMS are adopted, in terms of the responsibility for survey support operations. ... The Feasibility Study should consider the recommendations made in the mid-term reviews in relation to the role of PI teams in programming the survey and conducting real-time checks on data quality.

As discussed in more detailed below, it is not clear that quality will be improved if PI teams remove the responsibility for CAI programming and quality control from fieldwork agencies. The reviews also argue, however, that adopting these recommendations is a requirement for increasing competition between the agencies. Apropos of CAI programming, they argue:

With this change in work scope division, the bidding process for the field work would be more competitive because the work scope would be far more transparent with less exposure to unpleasant surprises as well as being within the grasp of any competent field organization regardless of its experience with longitudinal surveys. This change of responsibility should make the contracting-out of field work a more competitive process. Data collection contractors would resist this change as they want to hold on to the instrument programming task. This task, especially when it is executed using a proprietary system, protects them against competitors. Just as importantly, the specific human capital about the instrumentation is embedded in the contractor, functioning as yet another barrier against competitors. While the contractors might not like such a change, they would rather have 90% of the loaf than 0% of the loaf. In addition, there are other private sector firms that would be viable competitors except that they have no track record or expertise in complex longitudinal questionnaire software. (ULSC Report, pp 8-9; CLS Report, pp 20-21)

In this chapter we discuss whether and how competition and quality might be improved if PI teams take over responsibility for CAI programming and quality control, and we then consider practical steps that could be taken to improve competition if CAI programming and quality control remain the responsibility of the fieldwork agency. However it is important to start by outlining the level of competition that exists for national longitudinal studies in Britain compared with other countries.

Evidence suggests that competition for cohort and panel data collection contracts is perhaps more lively in the UK than in other parts of the world, with at between four and six major agencies that can be expected to bid for contracts. In the UK Ipsos MORI, NatCen, GfK NOP and TNS-BMRB are all capable of UK-wide face-to-face surveys, either alone or in consortium, as would be ONS and mruk.<sup>64</sup> In addition, while it is beyond the brief of this study to collect detailed comparative data, British cohort and panel studies are held in high international esteem and the costs involved compare favourably with other similar studies.

All the large scale longitudinal studies from which we have collected evidence report long-term collaborative relationships with fieldwork agencies and little competition for fieldwork contracts. CHRR, for example, has developed a system which is used as a 'turn-key'

---

<sup>64</sup> Ipsos MORI (<http://www.ipsos-mori.com/>); NatCen (<http://www.natcen.ac.uk/>); GfK NOP (<http://www.gfknop.com/>); TNS-BMRB (<http://www.tns-bmrb.co.uk/>); ONS (<http://www.ons.gov.uk/>); mruk (<http://www.mruk.co.uk/>). We note that the number of potential UK suppliers has recently been reduced, with the merger of TNS and BMRB and the creation of TNS-BMRB from their social sections.

solution with a variety of fieldwork providers. However, it has been running the NLS surveys, comparable in scope and complexity to those managed by CLS/ULSC, in partnership with NORC for more than two decades. Similarly, the other comparable studies from which we have information and which contract data collection on a competitive basis report between 1 and 3 individual or consortium bidders for the last contract commissioned; where successful contractors tend to change, the alternation is within a restricted pool of suppliers.

In the US, our understanding is that NORC, RTI and Westat are the major suppliers. In Europe, for example, the highly respected and oldest European panel study, the German Social Economic Panel, has never had more than two bidders, TNS-Infratest and Infas, with TNS-Infratest winning all contracts across the 25 year history of the study.<sup>65</sup>

### ***Responsibility for quality***

As we previously emphasised, (p11ff) quality control is not a zero-sum game; it is, and should be a responsibility shared between PI teams and fieldwork agencies. Removal of this function from the fieldwork agencies would mean that PI teams would lose access to the knowledge and expertise that agencies have built up over time, and on which their well-deserved reputations rest. Survey quality could therefore be put at risk by a strategy that removes quality control responsibilities from fieldwork agencies. It would also appear to be counter to one of the principal aims of the Survey Resources Network, i.e.:

... to improve the interface between academic Principal Investigators and professional survey organisations ... It is recognised that the majority of both expertise and excellence in UK social science survey research is located in professional survey organisations and that an enduring link between the professional survey world and the UK academic community is required to help build the survey research methods base within the academia. (TOR, 2.1)

Rather, ESRC should positively value the characteristics of the agencies that regularly tender for data collection contracts. All agencies with whom we have spoken, for example, have stressed the importance of quality control mechanisms, evidenced in the UK by the high level of certification to the international standard for market, opinion and social research (ISO20252).<sup>67</sup> ESRC could acknowledge the importance of this more explicitly by, for example, requiring compliance with ISO20252 and/or a detailed description of quality procedures, perhaps as a preliminary stage in the shortlisting procedures.

### ***Responsibility for CAI programming<sup>68</sup>***

Contrary to the hypothesis of the mid-term reviews, no survey agency with which we have communicated has expressed serious resistance to the idea of running CAI scripts which they have not written on systems which they do not themselves support. In fact, as many have pointed out, this situation is quite common in the UK industry where different agencies often act in consortia. In such situations, however, as was also made clear, the costs of licensing and support fall on the organisation providing the script. In the current situation, this would be the PI teams, and would represent an additional cost not hitherto borne.

We concluded above, however, that there is no reason in principle to recommend that the PI teams necessarily take on the CAI scripting role. Nor do we consider that there is any reason in principle to insist on the PI teams incurring additional costs by imposing specific CAI software on agencies that is different from that which the agencies already support. Logically, the only reason why the PI teams would insist on this is that the required CAI software is an integral and inextricable part of their overall data delivery systems. We have already questioned the value of closely-coupled system architectures, as they tend to be

---

<sup>65</sup> NORC (<http://www.norc.org/>); RTI (<http://www.rti.org/>); Westat (<http://www.westat.com/>); TNS-Infratest, <http://www.tns-infratest.com/>; Infas (<http://www.infas.de/home.html?&L=1&cHash=5ff413b71f>).

<sup>67</sup> Information from Bill Blythe, chair of the ISO 20252 technical committee.

<sup>68</sup> See also the section on Questionnaire Programming, above p7ff.

less flexible than modular-based development models. Equally, should these be 'off-the-shelf' systems, the PI teams would tend to be locked into a single supplier, as the more comprehensive the system, the greater the initial costs of implementation and subsequent costs of change. Further, having acknowledged the value of RDBMS, the only system based on RDBMS of which we are aware that could provide this comprehensive functionality is CHRR's Survey Suite. While we acknowledge that Survey Suite is a well-designed and extremely powerful product, estimates from CHRR are that it would take at least 3 years before it would be sufficiently portable for use by CLS/ULSC. Nor does this single-supplier solution rest easily with the aim of increasing competition between data collection agencies.

More specifically, we do not consider that 'the specific human capital about the instrumentation [that] is embedded in the contractor' is the major 'barrier against competitors' when it comes to surveys such as those managed by CLS/ULSC. Although both organisations aim to increase their use of phone and web interviewing, face-to-face interviewing will remain a core aspect of cohort and panel data collection. The collection of biological samples will typically require face-to-face interviewing and also restricts the choice of fieldwork agencies to those which can provide suitably qualified interviewers, often research nurses. The infrastructure required to do so is complex and the costs involved are not trivial. These would fall on the PI teams should the potential field of data collection agencies be extended without regard to their ability to meet this requirement, either individually or in consortium.<sup>69</sup> It is this core demand of the cohort and panel studies that we consider fundamentally limits the potential for competition between fieldwork agencies.

### ***Making the bidding process for fieldwork more competitive***

Limited competition does not, however, mean the absence thereof and, within these limits, ESRC should seek to maximise the benefits that it can bring. While we are not in a position to recommend any single course of action, we offer the following as suggestions ESRC should consider and, in conjunction with all stakeholders, elaborate and extend.

Most generally, competition between potential suppliers should not be viewed simply as a means of reducing costs between profit-seeking organisations. Costs must obviously be controlled and potential competitors seek a profit (or, at least, no loss) on investment. ESRC nevertheless must ensure that the quality and reputation of the product is maintained, and potential data suppliers seek other than purely financial benefits when bidding for contracts, however indirectly beneficial those benefits may be to their annual accounts.

As regards the latter, the cohort and panel studies, as well as other large-scale strategic data resources funded by ESRC, are already very attractive propositions. The agencies with whom we have consulted do not necessarily view them as large-profit vehicles, but the contracts on offer do provide a regular and substantial contribution to cash-flow. Equally, their scale and academic provenance make them extremely prestigious projects, which enhance the reputations of the successful agencies and contribute to their ability

---

<sup>69</sup> At the very least, the PI teams would have to manage relationships between multiple fieldwork suppliers. Additionally, as is the case at CHRR, the PI teams would also be responsible for the capital and recurrent costs of maintaining and distributing mobile computers. The mid-term reports consider these overheads to be minimal: "When the system is built to be Web-enabled, multi-modal surveys doing data capture over the Web (including cell phone internet connections), CATI or CAPI become simple to execute requiring only an appropriate Web browser ... CAPI is done either by putting a client and server on the laptop or tapping into the cellular network with a wireless modem and using the Web." (ULSC Report, 11; CLS Report, 23) The former solution ('putting a client and server on the laptop') is the one adopted by CHRR, who have extensively integrated the web and related technology into its systems. Notwithstanding this, CHRR supports an extensive infrastructure to maintain and communicate with these laptops, not including the overheads borne by NORC to integrate the software into its sample management systems. Re the latter solution, ("tapping into the cellular network"), we understand that the costs of mobile internet access are not yet sufficiently low in the UK to make this an economically viable method of conducting CAPI interviews of the kind deployed by CLS/ULSC. (cf Pazurick and Cameron (2007, 44): "It is impossible to predict the future, but the disruptive effect of mobile technology on survey research seems inevitable. What remains to be seen is exactly how that disruption will unfold and how long it will take to do so.")

successfully to bid for other projects. As one fieldwork agency representative told us, the cohorts and panels are studies for which 'it is impossible not to bid.'

Having said this, we acknowledge that the 'bottom line' is as important to potential contractors as it is to ESRC, and while we expect that ESRC will not seek to fund a survey at a cost below which it is financially viable for fieldwork agencies, we also do not expect ESRC to provide infinite funds to enhance inducements to bid. However, ESRC could attempt to reduce the costs of bidding by, for example, funding the process for short-listed agencies or by simplifying the bidding process, e.g. by providing a detailed, but formulaic structure in which bids could be presented. This could also be used as a transparent 'score-card' used to structure the evaluation of competing bids and to provide feedback to unsuccessful bidders. Additionally, ESRC could consider the value of offering single tender contracts subsequent to, and conditional on performance on, an initial contract won on a competitive basis. This could help agencies better plan the amortisation of any significant startup costs and provide an extra performance incentive.

ESRC could also enhance the attractiveness of a contract by working with PI teams to ensure that the invitation to tender (ITT) is sufficiently detailed to enable the bidders fully to understand the requirements. The ITT could potentially also state the maximum funds available so that incumbent providers do not have the competitive advantage of knowledge about likely levels of funding based on previous experience. All agencies with whom we have consulted suggested that fixed-cost contracts are the norm by virtue of client preference. The benefits to clients are obvious, as a great deal of risk is transferred to the data collection agency, but the agencies do, in consequence, require sufficient detail in the ITT to assess whether or not bearing the risk is viable. Similarly, it has been suggested that awarded contracts should be worded in sufficient detail to ensure that the scope of work requirements do not significantly exceed the expectations engendered by the ITT – while it is acknowledged that project needs cannot be completely determined in advance, excessive 'requirements creep' can have negative impacts in the long term, e.g. increasing the margins that agencies build in when bidding for subsequent contracts and/or reducing their incentive to bid.<sup>70</sup>

In addition to augmenting the attraction of a bid, ESRC in conjunction with PI teams, could also research potential suppliers and ensure that they are directly informed of the ITT, thus ensuring that the pool of bidders is maximised.

Regardless of how attractive a data collection contract might seem in the abstract, potential bidders will only enter the contest if there is at least the possibility of a tangible return. However, the probabilities of success for the long term studies managed by CLS/ULSC tend, as a matter of fact, to reduce over time. NatCen, for example, has been almost exclusively the contractor for the cohort studies over the past decade, while all 18 waves of BHPS were conducted by NOP GfK, a continuity reflected elsewhere.<sup>71</sup>

This situation is recognised by both PI teams and fieldwork agencies alike, and is only to be expected by virtue of the structural inertia that increases with the length of the study. The initial rounds of a long-term longitudinal study that involve a new data collection agency are almost by definition beset by problems stemming from the novelty of the exercise and, hence, unpredicted requirements, as well as the efforts on the parts of both the PI team and the data collection agency to establish an effective working relationship. Once these initial problems are overcome, the scientific team's subsequent assessment of the risks of changing agencies will tend to emphasise conservatism, rooted in the principle of 'if it ain't broke don't fix it'. This conservatism is reinforced by the positive value of maintaining

---

<sup>70</sup> We note that agencies are not reticent to charge more for requirements clearly not included in the original contract or scope of work. The problem we allude to here is of the grey area created by loosely specified requirements which escalate in scope and thus cost as they are progressively refined.

<sup>71</sup> TNS-Infratest has been even more successful with the German Socio-Economic Panel, having been the data provider for all 25 waves of the study.

interviewer continuity, which contributes to sample retention. On the hand, once an agency has acquired a contract as prestigious as those for the cohorts or panel studies, its reputation would suffer should it lose the contract at a subsequent stage. In consequence, incumbent agencies are, and are perceived to be, highly motivated to maximise performance and the competitive attractiveness of their bids in subsequent rounds. This positive effect is mitigated to the extent that both the incumbent agency and the competition see little possibility of a change in contractor.

ESRC should therefore risk-assess contracts with data collection agencies to ensure that the incumbent agency does not achieve any insurmountable advantage in bids for subsequent contracts.<sup>72</sup> This could involve, for example, a contractual obligation on the PI team to prepare an annually reviewable 'survey continuity plan' in which the risks of change in data supplier are explicitly addressed – whether this be due to the invocation of a 'break' clause due to poor performance or a change of supplier due to the award of a new contract – as well an obligation on the data collection agency to contribute to this plan with its own plans for transfer of responsibilities to another agency.

One suggestion put to us was that PI teams should not be part of the panel evaluating bids for and awarding fieldwork contracts. This, it is argued, would eliminate any bias or perception of bias on the part of PI teams towards particular fieldwork agencies. We agree that the evaluation panel should include members independent of the PI teams to ensure that all bids are objectively evaluated, but also consider that PI teams must be represented on the evaluation panel as well given their specific expertise in study requirements.

We accept the value of competition between data collection agencies and consider that ESRC can examine a number of ways to increase it. We caution strongly against a strategy that attempts to achieve this objective by removing responsibility for quality and for CAI programming from the competitors in order to increase their number, as this would risk survey quality and possibly increase costs. Nor have we found evidence to support the likelihood of success or desirability of this outcome.

*Recommendation 19. ESRC consult with stakeholders to determine the most appropriate means of augmenting competition for data collection agencies within a limited market place by enhancing the prestige of the studies and valuation of suitable agencies, lowering the costs of bidding for, and enhancing the detail of Invitations to Tender, and minimising the inherent advantages of an incumbent contractor.*

## **Chapter 6. Conclusions. Towards a Common Infrastructure**

We began our research for this Report in agreement with the basic assumption of the mid-term review reports that the effective use of metadata throughout the survey process could offer both efficiency and quality gains, an assumption which our research has reinforced and which lays behind a number of our recommendations.<sup>73</sup> These gains stem from the fact that metadata-driven systems reduce redundancy in at least two ways. First, information can be reused throughout the process and thus needs to be captured only once. Secondly, systems can be comprised of generic components that are more robust and adaptable to varying process requirements. However, our research also lead us to conclude that quality and efficiency across the survey process as a whole is determined by many other factors, and that the relationship between them is less straightforward than assumed by the mid-term review reports and reflected in the TOR. For example, fieldwork agencies do not currently have sole responsibility for data quality control in the cohort or panel studies and nor, by contrast, should CLS/ULSC themselves seek sole responsibility. Nor, given the operating environment, do we consider that the CLS/ULSC can or should necessarily take on the role

---

<sup>72</sup> In this respect, we note that the current USoc contract is based on survey waves which have overlapping interview cycles. Any change of data collection agency at the end of a contract period will thus result in ULSC simultaneously interfacing with two agencies, consequently increasing the risks of contractor change.

<sup>73</sup> See Recommendation 2, p7, Recommendation 11, p14 and Recommendation 12, p15.

of programming CAI instruments. Equally, we have concluded that there are limits to the number of fieldwork agencies capable of collecting data for the cohort or panel studies and that any increase in competition between them will be a function of factors other than the locus of data quality control or instrument programming. However it is also clear that competition for fieldwork does exist in Britain and at a higher level than in several other countries.

Our investigations have aimed to cover the entirety of the survey process or 'data production line', and this is reflected in the conclusions we have drawn and the recommendations we have made. Within the confines of this study, it has not been possible to prioritise the work implied by our recommendations and to present a coherent development plan through which they can be realised. We are therefore limited in the extent to which we can:

...examine the use of web-enabled tools ... to offer further cost efficiencies through a common e-infrastructure shared between the Centres ... The Study should investigate the infrastructure required for these operations [CAI programming and real-time quality checks] to be carried out by the PI teams, both in terms of upgraded physical/computing resources and additional staffing. The Study should then suggest a forward look strategy to enable the development of appropriate infrastructure and resources at the two Centres, including likely costs. (3.3, 3.6)

Development of a strategic plan requires detailed cost-benefit analyses, comparing the work involved in realising our recommendations against the expected outcomes that could be achieved, and should take into account the considerations discussed below.

It should include an initial development cycle that concentrates at the outset on modelling the metadata required to support the survey process and, in particular, on building or acquiring tools for metadata capture at the point of questionnaire specification. The development programme should align to the greatest extent possible with the timetable for the new 2012 cohort and must take into account the on-going and different requirements of existing surveys. Operational deployment within these environments should therefore be expected to be evolutionary, not necessarily following the same pattern in each, with development based on a modular strategy. Systems development should use third party components and applications as far as possible, using 'web-enabled tools' (TOR, 3.3)<sup>74</sup> wherever possible. Open-source software should also be used wherever possible, and all appropriate software developments should also be made available under an open-source licence – as publicly funded bodies, CLS/ULSC have an obligation to make use of and contribute to the open-source stockpile. We do not recommend, however, considering a complete 'off the shelf' solution (TOR, 3.4). As the mid-term review reports notes, these solutions do not necessarily support the requirements of CLS/ULSC surveys (ULSC Report, 12; CLS Report, 24). More generally, a complete "all-in-one" solution would force CLS/ULSC into operational practices dictated by the system, rather than survey requirements. A modular strategy, building on the development of relatively independent components interacting through standard interfaces, will produce more flexible and robust systems, better able to meet both the common and individual needs of the different surveys both as they are and as they develop in the future. Systems development should maximise returns on existing investments. On the one hand, existing applications and subsystems should be replaced as required, as implied by the evolutionary deployment model suggested above. For example, even if SIR™ is not used as the strategic data repository, until they can be replaced, existing PQL™ based QC and value-adding systems could continue to be used. On the other, existing CLS/ULSC applications and system components may be able to be incorporated 'as-is' or, at least, provide the basis on which more suitable products can be developed (e.g. the Understanding Society Portal or questionnaire specification tools currently being developed by ULSC). Given the extreme sensitivity of the data managed by CLS/ULSC, security must be of paramount importance in all systems development.

---

<sup>74</sup> For example, using browser based interfaces and web services-based architectures (W3C, 2004).

Information security management systems should therefore be based on ISO/27001, a recognised international standard for information security management. (ISO27001)

The TOR suggests two different models for the implementation of a 'common e-infrastructure': on the one hand, a single facility, shared by CLS/ULSC but maintained independently of both or, on the other, a set of common technologies installed and maintained separately by each organisation. The models are not necessarily mutually exclusive. Nor does either model have overwhelming *prima facie* advantages over the other. A central facility, for example, could provide basic infrastructural services such as secure data storage and web servers, with analytic or value added processing offloaded to local environments in which desktop tools might prove more efficient. It could also potentially offer economies of scale and be available for use by other ESRC assets. It would, however, require the creation of a new entity, with additional attendant initial and ongoing administrative costs, appropriate arrangements to ensure that CLS/ULSC continue to meet their confidentiality obligations to respondents, and detailed service level agreements to define and meet delivery needs. Distributed installation of common technologies would avoid these additional costs and could leverage existing infrastructure support, but could make it more difficult for other ESRC assets to avail themselves of their benefits, and could also reduce CLS/ULSC motivation to adopt new and common technologies in favour of persisting with existing and familiar tools. Similar considerations apply to the allocation of staff resources required to develop and maintain the operational systems themselves.

The costs of establishing a common e-infrastructure and consequent effects on the staff profiles and the Centres are thus impossible to state with any precision at this time, though we will send ESRC indicative costs under separate cover. To achieve a strategic plan in which costs and effects can be more definitely quantified, ESRC must further engage with all stakeholders. As a starting point, ESRC should promote nascent forms of co-operation between the Centres, actively seek widespread views on our analyses and recommendations.

One, perhaps unintended, consequence of producing this Report has been the establishment of closer working relationships between CLS and ULSC, the result of which has been preliminary discussions as to how progress might be made on some recommendations as part of on-going development activities, in particular with respect to the questionnaire specification tools on which ULSC has already been working. (see footnote 18, p7) Similarly, we are pleased to note that we have constructively engaged many of the individuals from whom we were seeking the information that has informed this Report. As more than one informant remarked, few fora exist to discuss in technical detail the questions addressed by this Report.<sup>76</sup> ESRC should also host at least one conference, and perhaps a regular series of events, focussed on the issues raised and recommendations made by this Report both to facilitate the development of a strategy to implement our recommendations.

*Recommendation 20. ESRC should engage with stakeholders to prepare a strategic plan for the realisation of a common e-infrastructure. The plan should provide a model of how the infrastructure will be implemented and how development resources will be allocated and prioritised. ESRC should promote co-operative work between the Centres on components that could form part of a common infrastructure.*

*Recommendation 21. ESRC should actively seek widespread views on this Report. As part of this process, ESRC should organise at least one conference on the topics covered by this Report.*

We conclude by noting that the TOR focussed on the needs of the cohort and panel studies as conducted by CLS and ULSC; in consequence, we have also done so in this Report. Very few of our analyses or recommendations are necessarily restricted to either organisation or

---

<sup>76</sup> One intentional exception to this rule is the Association for Survey Computing. (<http://www.asc.org.uk>)

the studies for which they are responsible, and ESRC should view this Report as being of potential relevance to the larger set of survey resources that it funds.

## References

- CLS Report *Report of the Mid Term Review for the Research Resources Board of the Economic and Social Research Council into the Centre for Longitudinal Studies (CLS)*, November 2007. (unpublished)
- Calderwood, Lisa (2009). *The relationship between survey quality and survey processes in longitudinal studies*. CLS Working Paper (forthcoming) [Copy available on request]
- Codd, EF (1970). A relational model of data for large shared data banks. *Communications of the ACM* 13 (6), pp 377-387. (available from <http://portal.acm.org/citation.cfm?id=358007>)
- Costigan, Paddy and Elder, Steve (2003) "Does the Questionnaire Implement the Specification? Who Knows?" in Banks, R *et al* (eds) ASC2003. The Impact of Technology on the Survey Process. Proceedings of the Fourth International Conference on Survey and Statistical Computing, presented by The Association for Survey Computing, 27-19 September, pp. 85-96.
- Couper, Mick (2007) "Whither the Web", in Trotman, Mike *et al* (eds) The Challenges of a Changing World. Proceedings of the Fifth International Conference of the Association for Survey Computing, 12-14 September 2007, pp7-16.
- DSD (2006). US Census Bureau. Demographic Surveys Division. SPIDER Users Guide. (last updated, 1 December 2006)
- Groves, Robert M. (1989) *Survey Errors and Survey Costs*. New York: John Wiley
- Robert M., and Steven Heeringa (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169, 439-457 Part 3. (available from <http://www3.interscience.wiley.com/cgi-bin/fulltext/118612304/PDFSTART>)
- Harold, Elliotte Rusty (2005). Managing XML data: Native XML databases. Dated 2005-06-06 at <http://www.ibm.com/developerworks/xml/library/x-mxd4.html>
- Harold, Elliotte Rusty (2007) The State of Native XML Databases. (<http://cafe.elharo.com/xml/the-state-of-native-xml-databases/>)
- Hill, David, Kass, Christina, Reed-Gillette, Debra, Berman, Lewis (2001). Use of Blaise™ in the National Health and Nutrition Examination Survey. Presented to the 2001 7th International Blaise Users Conference. ([http://www.blaiseusers.org/2001/papers/Hill-IBUC\\_Paper\\_final\\_NCHS\\_approved.pdf](http://www.blaiseusers.org/2001/papers/Hill-IBUC_Paper_final_NCHS_approved.pdf))
- ISO20252 ISO 20252:2006 Market, opinion and social research ([http://www.iso.org/iso/catalogue\\_detail?csnumber=39339](http://www.iso.org/iso/catalogue_detail?csnumber=39339))
- ISO27001 ISO/IEC 27001:2005. Information technology -- Security techniques -- Information security management systems – Requirements. ([http://www.iso.org/iso/iso\\_catalogue/catalogue\\_tc/catalogue\\_detail.htm?csnumber=42103](http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=42103))
- Lynn, Peter (2001). *A Quality Framework for Longitudinal Studies*. (Draft) Last update 26-9-2001. (available from <http://www.iser.essex.ac.uk/files/ulsc/standards/framework/2001-09-26.pdf>)
- Pazurick, Aaron and Cameron, Mark (2007) Survey Research in a Wireless World in Trotman, Mike *et al* (eds) *The Challenges of a Changing World*. Proceedings of the Fifth International Conference of the Association for Survey Computing, 12-14 September 2007, pp39-46. SRC (2008) – SRC Blaise Standards. 2008\_1 Edition (March, 2008)
- Philology (2004). QEDML Technology Overview. A presentation to Automation in the Survey Process. Managing Change and Avoiding Disaster, An ASC Software Showcase,

- Thursday 22nd April 2004 (available from  
<http://www.asc.org.uk/Events/Apr04/Contributions/gedml.pdf>)
- SRC (2008) – SRC Blaise Standards. 2008\_1 Edition (March, 2008)
- Statistics Canada (1998) *Statistical Quality Guidelines* (3<sup>rd</sup> Edition). Ottawa: Statistics Canada (available from <http://www.statcan.gc.ca/pub/12-539-x/4194542-eng.pdf>)
- ULSC Report. *Report of the Consultant for the Research Resources Board of the Economic and Social Research Council into the UK Longitudinal Studies Centre (ULSC)*, November 2007. (unpublished)
- Williams, Ann (2005) Performance of relational databases versus native XML databases. Honours thesis, University of Otago (Information Science).  
(<http://eprints.otago.ac.nz/323/2/AnneWilliamsOCR.pdf>)
- W3C (2004). Web Services Architecture. W3C Working Group Note 11 February 2004.  
(<http://www.w3.org/TR/ws-arch/>)
- W3C (2008) Extensible Markup Language (XML) 1.0 W3C Recommendation 10-February-1998. (<http://www.w3.org/TR/1998/REC-xml-19980210>)

## **Appendix A. Terms of Reference**

The following are the terms of reference for this study, as issued by the ESRC in April 2008 to short-listed potential applicants for the Survey Resources Network.

### **Integrated Survey Data Collection, Fieldwork Management and Survey Data Processing Systems for Longitudinal Studies:**

#### **Terms of Reference for a Feasibility Study**

##### **1. Background**

- 1.1. The Economic and Social Research Council (ESRC) began the process of commissioning a Survey Resources Network in January 2008. The aim of the Survey Resources Network is to provide a better foundation for the Council's investments in the strategically important area of survey research, facilitating more efficient consolidation and co-ordination of these resources. The Survey Resources Network is to be affiliated to the National Centre for Research Methods (NCRM) for the purposes of information sharing and dissemination of relevant outputs.
- 1.2. The proposed Network has five main objectives:
  - (1) Fostering and promoting the development of new methods within survey methodology ('survey research')
  - (2) Providing high quality online resources that can be used for training and research within the area of survey research;
  - (3) Contributing to building capacity in high-quality survey practice;
  - (4) Coordinating all of the above activities at a national and international level;
  - (5) Establishing the feasibility, potential efficiency gains and quality improvements in the processes relating to:  
data collection (via different modes of data collection, the use of survey case management systems, and so on), sample maintenance, data coding and editing methods, and the preparation of metadata to common standards.
- 1.3. This document refers to Objective 5 and sets out Terms of Reference for the Feasibility Study to be carried out under this objective, with specific reference to longitudinal studies funded by the ESRC. As stated in the specification for the Survey Resources Network, this objective is borne out of the mid-term reviews of the Centre for Longitudinal Studies (CLS) and the UK Longitudinal Studies Centre (ULSC) conducted by Professor Randy Olsen.

##### **2. The Need**

- 2.1. One of the principal aims of the Survey Resources Network is to improve the interface between academic Principal Investigators and professional survey organisations. The health of survey-based social and economic research in the UK depends, in turn, on the health of the methods and techniques employed to generate survey data. It is recognised that the majority of both expertise and excellence in UK social science survey research is located in professional survey organisations and that an enduring link between the professional survey world and the UK academic community is required to help build the survey research methods base within the academia.
- 2.2. Mid-term reviews of both CLS and ULSC, conducted by an independent consultant, highlighted similar issues in terms of the relationship and balance of work between academic teams involved in large-scale longitudinal surveys and fieldwork agencies to whom large portions of work are currently outsourced. One of the central recommendations of the mid-term reviews was for the ESRC to review current data management processes and systems in place at the two Centres, including the degree to which survey support operations can be undertaken by academic PI teams, the way

in which fieldwork is procured and the use of relational database management systems.

- 2.3. In particular, it was recommended that the Centres should build up their infrastructure so that they are capable of handling many of the survey support operations they currently subcontract out to vendors. Three specific points were made in relation to this recommendation:
  - ULSC/CLS should program the instrument for delivery to the data collection subcontractor;
  - ULSC/CLS should receive the data from the field on a daily basis, load it into their database, and engage in real-time data quality and interview validity;
  - ULSC/CLS databases for the survey data should have a table schema of sufficient completeness to document key features of the survey specification as well as index and document the variables across a variety of dimensions to facilitate the search for variables and information about the variables by new users in the downstream data dissemination process.
- 2.4. Once the Centres have established themselves as being able to support these functions, it was recommended that they structure their procurement of fieldwork so that the sub-contractor's responsibilities focus on training and management of the fieldwork interviewers, as opposed to on supporting specialised technical infrastructure for programming complex longitudinal studies.
- 2.5. In addition, it was recommended that both Centres re-examine the use of the Scientific Information Retrieval system, commonly known as SIR, as the relational database management system used to support survey activities. The mid-term reviews highlight the developments made in relational database management systems since SIR was adopted in the 1980s and recommend that processes at CLS and ULSC would benefit from being updated with a modern relational database survey management system. In particular, it was recommended that the Centres collaborate to choose a product with a large market share (e.g. Oracle, Sybase, IBM or Microsoft) which could offer a range of opportunities, such as effective training of existing staff, hiring of new staff with experience in using the product, and adopting a range of add-on utilities, as well as seeking cost efficiencies between the CLS and ULSC. By taking a long term approach and structuring their infrastructure around relational database methods, it was suggested that the Centres could be more effective in integrating their survey support activities, as well as offering these support activities to smaller groups working on longitudinal projects.
- 2.6. As a result, the ESRC decided to use the opportunity presented by the commissioning of the Survey Resources Network to conduct a feasibility study to look at potential efficiencies in data management processes, particularly in relation to data management software and the use of cutting edge data collection methods for longitudinal surveys carried out at CLS and ULSC. This is addressed by Objective 5 in the specification for the Network.

### **3. Detailed Terms of Reference**

- 3.1. The Feasibility Study should consider two major issues highlighted in the CLS/ULSC mid-term reviews: the scope of work undertaken at the two Centres and the supporting technical survey infrastructure. These two issues are strongly linked. At the heart of the mid-term reviews of the Centres, is a suggestion that it would be beneficial for CLS and ULSC to have greater control over all the survey process from design through to data dissemination. This Feasibility Study should examine these arguments for change, so that the risks associated with – and with not – altering the work scope and supporting technical infrastructure are properly considered. The Study should then investigate what systems and processes should be put in place to help facilitate the

proposed changes and what impact these would have on the way in which the Centres are currently configured.

### **Survey Technology**

- 3.2. The mid-term reviews argue that the use of relational database management systems (RDBMS) holds the key to reforming the balance of work between the academic PI teams based at the Centres and the fieldwork agencies. It is acknowledged that the current SIR system is outdated and that modern RDBMS are the most effective tools for work with, and archiving, longitudinal survey data. The Feasibility Study should therefore investigate the variety of RDBMS available, in conjunction with the requirements of the longitudinal studies, to ascertain appropriate options for CLS and ULSC.
- 3.3. In particular, the Feasibility Study should evaluate how RDBMS can be used to offer improvements to:
  - a) survey design - through the development of linked questionnaire specifications;
  - b) data collection - through interaction with the underlying database structure;
  - c) data dissemination - through links to the linked questionnaire specifications and linked information that helps document the individual questions and their associated data being made available to the UK Data Archive.

In all these elements, the Feasibility Study should examine the use of web-enabled tools which can interact with RDBMS to offer further cost efficiencies through a common e-infrastructure shared between the Centres.

- 3.4. The Feasibility Study should investigate the pros and cons of 'off the shelf' RDBMS used for cross-sectional and market research purposes versus bespoke products. In particular, the Study should consider mechanisms for incorporating legacy data from previous waves of the longitudinal studies. In assessing the variety of RDBMS available, the Study should take account of international best practice in the operation of longitudinal studies elsewhere in the world. The Study should also be mindful of the needs of UK longitudinal studies looking forward, such as the new UK Household Longitudinal Study (UKHLS) and the proposed new birth cohort study in 2012.

### **Work Scope**

- 3.5. Alongside technical infrastructures, the Feasibility Study should examine the future work scope of CLS and ULSC in terms of their interaction with external partners. In particular, the Study is asked to consider the relationship between the academic PI team and fieldwork agency, if RDBMS are adopted, in terms of the responsibility for survey support operations.
- 3.6. The Feasibility Study should consider the recommendations made in the mid-term reviews in relation to the role of PI teams in programming the survey and conducting real-time checks on data quality. The Study should also consider the work of the PI teams in preparing data for deposit at the UK Data Archive and potential efficiencies which could be gained through preparing metadata to common standards. The Study should investigate the infrastructure required for these operations to be carried out by the PI teams, both in terms of upgraded physical/computing resources and additional staffing. The Study should then suggest a forward look strategy to enable the development of appropriate infrastructure and resources at the two Centres, including likely costs.

## **4. Available Resources**

- 4.1. In taking looking at the various options in relation to survey technology and work scope, the Consultant undertaking the Feasibility Study will be expected to liaise with CLS and ULSC, as well as representatives of the ESRC and co-funders of the longitudinal studies. A working group has been set up by the ESRC to take forward the

recommendations of the CLS/ULSC mid-term reviews and the Consultant will be expected to report progress to the group.

4.2. The following documents will be made available to the Consultant:

- a) Submissions to the mid-term review process by CLS and ULSC;
- b) Mid-term review reports for CLS and ULSC by Professor Olsen;
- c) Responses to the mid-term reviews by CLS and ULSC.

## 5. Timetable

5.1. The key dates of the commissioning process are set out below:

Date/Period	Event
26 February 2008	Deadline for outline proposals to be submitted
End February – Mid April 2008	Assessment of outline proposals
End April 2008	Decision letters sent to outline applicants
27 May 2008	Deadline for full proposals for Survey Resources Network to be submitted
End May – Mid-August 2008	Assessment of full Survey Resources Network proposals
Late August 2008	Decision letters issued
1 November 2008	Survey Resources Network and Scoping Study commences

5.2. At this stage, we anticipate that the Feasibility Study should be completed and submitted to the ESRC Research Resources Board by 1 March 2009<sup>77</sup>. We anticipate the final report will be no more than 25 A4 pages in length.

---

<sup>77</sup> As noted earlier, this deadline was subsequently altered to 31 October 2009.

## Appendix B. Recommendations

- Recommendation 1. CLS/ULSC should investigate the use of web-based discussion fora and issue management software, as well as, more generally, of appropriate 'Web 2.0' technologies, in the process of consultation with the wider scientific community..... 6*
- Recommendation 2. CLS/ULSC should adopt and define where required common questionnaire specification standards that are independent of any particular CAI system. These standards should include the metadata models that describe questionnaires and software tools to repurpose and use the information collected during the specification process. The standards adopted should align with the open-standard DDI metadata model in order to leverage emergent tools and facilitate data documentation, dissemination and archiving. .... 7*
- Recommendation 3. Responsibility for programming the CAI instruments should be allocated after CLS/ULSC determine with their fieldwork contractors how best to achieve the highest quality instruments through the most efficient development processes. .... 9*
- Recommendation 4. CLS/ULSC should contribute to the development of new, and the improvement of existing, methods to test CAI instruments and should use current best practice for testing. The development schedule should provide sufficient time for iterative testing of instruments prior to deployment in the field. .... 9*
- Recommendation 5. CLS/ULSC should investigate design options to enable the development of an extensible, general purpose sample management system (SMS) to support both the common and unique needs of cohort and panel surveys. SMS requirements should be established by analysing the SMS currently used by CLS/ULSC, as well as those used by other organisations, to ensure maximum generalisability..... 10*
- Recommendation 6. Fieldwork agencies should provide CLS/ULSC with daily fieldwork monitoring information either as separate data feed or via the systems used by agency staff themselves subject to agreements about the use and distribution of this information. .... 10*
- Recommendation 7. CLS/ULSC should review existing interviewer training, allocation, payment structures and fieldwork quality control procedures and consider whether changes could lead to cost effective improvement in survey quality. .... 11*
- Recommendation 8. Responsibility for data quality should be shared between PI teams and fieldwork agencies in order to achieve the high standards required by the cohort and panel surveys..... 12*
- Recommendation 9. ESRC should require costed details of data quality control systems as part of tenders for strategic data resources to which CLS/ULSC and others might respond. At least one referee for the bids should possess the technical expertise required to evaluate the proposals..... 12*
- Recommendation 10. Fieldwork contractors should provide CLS/ULSC with daily data and paradata delivery. CLS/ULSC should develop quality-control and data throughput systems which use this data on a continuous basis. .... 13*
- Recommendation 11. CLS/ULSC should adopt common data output specification standards that are CAI- and data management system-independent. These standards should be used to specify the content, structure and format of all data to be delivered. Specifications should be generated from metadata collected at the questionnaire specification stage to allow data delivery requirements to inform questionnaire programming and be tested simultaneously, well in advance of the commencement of fieldwork. Data management staff should be directly involved in questionnaire*

specification to ensure that the logic of the data collection instrument aligns with output requirements. ....	14
<i>Recommendation 12. CLS/ULSC should develop a generic metadata model on which a common repository can be developed to support survey processing systems. This model should be extensible and implementation independent, incorporate both the common and specific features of the surveys they currently manage, and align with other relevant metadata standards. The development process should investigate and, as possible and appropriate, use existing metadata models supporting comparable projects.....</i>	15
<i>Recommendation 13. CLS/ULSC should work closely with the UKDA to achieve DDI-compliant metadata deposit.....</i>	16
<i>Recommendation 14. ESRC should fund at least one of CLS/ULSC to become a member of the DDI Alliance, specifically to contribute developments of benefit to longitudinal surveys in general, and the cohort and panel surveys in particular. ....</i>	16
<i>Recommendation 15. CLS/ULSC should establish common standards for distribution and documentation of longitudinal data based on best international practice.....</i>	17
<i>Recommendation 16. CLS/ULSC should work with the UKDA to explore options for data dissemination that provide both centralised user authentication and distributed access to the data and documentation through systems specifically designed to meet the needs of data users. ....</i>	17
<i>Recommendation 17. CLS/ULSC should continue to develop existing mechanisms for user training, and should make significant investments in creating web-based training resources following best international practice.....</i>	18
<i>Recommendation 18. In selecting an RDBMS to replace SIR and extend the use of RDBMS technology, CLS/ULSC should prioritise platform-independent, open-source RDBMS that provide good support for the SQL™ standard. CLS/ULSC should take into account their use of existing general purpose RDBMS platforms, and the extent to which their use is application dependent and/or can be generalised. CLS/ULSC should also investigate the extent to which RDBMS technologies can be replaced with XML and XML databases or XML-enabled RDBMS.....</i>	20
<i>Recommendation 19. ESRC consult with stakeholders to determine the most appropriate means of augmenting competition for data collection agencies within a limited market place by enhancing the prestige of the studies and valuation of suitable agencies, lowering the costs of bidding for, and enhancing the detail of Invitations to Tender, and minimising the inherent advantages of an incumbent contractor.....</i>	25
<i>Recommendation 20. ESRC should engage with stakeholders to prepare a strategic plan for the realisation of a common e-infrastructure. The plan should provide a model of how the infrastructure will be implemented and how development resources will be allocated and prioritised. ESRC should promote co-operative work between the Centres on components that could form part of a common infrastructure.....</i>	27
<i>Recommendation 21. ESRC should actively seek widespread views on this Report. As part of this process, ESRC should organise at least one conference on the topics covered by this Report.....</i>	27

## Appendix C. List of Acronyms

BCS	British Cohort Study (1970 Birth Cohort)
BHPS	British Household Panel Survey
CAI	Computer-assisted interview
CAPI	Computer-assisted personal interview
CATI	Computer-assisted telephone interview
CHRR	The Center for Human Resource Research (Ohio State University)
CLS	Centre for Longitudinal Studies (Institute of Education)
DBMS	Database management system
DDI	Data Documentation Initiative
DPL	Data production line
ECLS	The Early Childhood Longitudinal Study (NCES)
ESRC	Economic and Social Research Council
HRS	Health and Retirement Study (University of Michigan)
ICPSR	The Inter-University Consortium for Political and Social Research (University of Michigan)
ISER	Institute for Social and Economic Research
ITT	Invitation to tender
MCS	Millennium Cohort Study (200-01 Birth Cohort)
MESS	Measurement and Experimentation in the Social Sciences
NatCen	The National Centre for Survey Research (London)
NCDS	National Child Development Study (1958 Birth Cohort)
NCES	National Center for Education Statistics (Washington, DC)
NCHS	National Center for Health Statistics (Hyattsville, Maryland)
NHANES	The National Health and Nutrition Examination Survey (NCHS)
NLS	National Longitudinal Survey
NORC	The National Opinion Research Center (University of Chicago)
ONS	Office for National Statistics
PI	Principal Investigator
PQL	Procedural Query Language
PSID	Panel Study of Income Dynamics (University of Michigan)
QC	Quality control
QDT	Questionnaire Development Tool
QEDML	Questionnaire Exchange and Deployment Markup Language
RDBMS	Relational database management systems
SAS	Statistical Analysis System (software)
SDMX	Statistical Data and Metadata Exchange
SIR	Scientific Information Retrieval
SRC	Survey Research Center (University of Michigan)
SOEP	Socio-Economic Panel (Berlin)
SPIDER	SPecialized Instrument DEvelopment Resource (US Census Bureau)
SQL	Structured Query Language
SRC	The Survey Research Center (University of Michigan)
SRN	Survey Resources Network
SPSS	Statistical Package for the Social Sciences (software)
TOR	Terms of Reference
UKDA	UK Data Archive (University of Essex)
UKHLS	Understanding Society: the UK Household Longitudinal Study
ULSC	United Kingdom Longitudinal Studies Centre
USoc	Understanding Society (the UK Household Longitudinal Study)
VPN	Virtual Private Network

## Appendix D. Details of Meetings Held

Date	Visiting	Attendees
15 Jan 2009	CLS/ULSC, University of Essex ( <a href="http://www.cls.ioe.ac.uk">http://www.cls.ioe.ac.uk</a> and <a href="http://www.iser.essex.ac.uk/ulsc">http://www.iser.essex.ac.uk/ulsc</a> )	Heather Laurie, Fran Williams, Elaine Prentice-Lane (ULSC); Jane Elliott, Mac McDonald, Lisa Calderwood, Jon Johnson, Randy Banks, Geoff Angel
26 Jan 2009	NORC, University of Chicago ( <a href="http://www.norc.uchicago.edu">http://www.norc.uchicago.edu</a> )	Kymn M. Kochanek, Vickie Wilmer, Dan Black (NORC); Jane Elliott
29 Feb 2009	UKDA, University of Essex ( <a href="http://www.data-archive.ac.uk/">http://www.data-archive.ac.uk/</a> )	Matthew Woollard, Ken Miller (UDKA); Randy Banks, Geoff Angel
30 Mar 2009	SRC, University of Michigan ( <a href="http://www.src.isr.umich.edu">http://www.src.isr.umich.edu</a> )	Bob Groves (SRC); Jane Elliott, Lisa Calderwood, Jon Johnson, Randy Banks
30 Mar 2009	ICPSR, University of Michigan ( <a href="http://www.icpsr.umich.edu">http://www.icpsr.umich.edu</a> )	Jared Lyle, George Alter, Peter Granda, David Thomas, Cole Whiteman, Mary Vardigan, Sue Ellen Hansen, Sanda Ionescu, James McNally (ICPSR); Jane Elliott, Lisa Calderwood, Jon Johnson, Randy Banks
31 Mar 2009	PSID, University of Michigan ( <a href="http://src.isr.umich.edu/src/psid">http://src.isr.umich.edu/src/psid</a> )	Frank Stafford, Bob Schoeni, Kate McGonagle, Eva Lessiou, Mohammad Mushtaq (PSID); Jane Elliott, Lisa Calderwood, Jon Johnson, Randy Banks
31 Mar 2009	HRS, University of Michigan ( <a href="http://hrsonline.isr.umich.edu">http://hrsonline.isr.umich.edu</a> )	Kathy Terrazas, Theresa Norgard (HRS); Jane Elliott, Lisa Calderwood, Jon Johnson, Randy Banks
1-2 Apr 2009	CHRR, University of Ohio ( <a href="http://www.chrr.ohio-state.edu">http://www.chrr.ohio-state.edu</a> )	Randy Olsen, Karima Nagi, Canada Keck, Margaret Lowden, Frank Marino, Brent Bogreese, Nick Ramser, Rosella Gardecki, Craig Lee, Jon Berry, Nicki Cartt, Ji Zhang, Jaron Shook (CHRR); Jane Elliott, Lisa Calderwood, Jon Johnson, Randy Banks, Geoff Angel
6 Apr 2009	NORC, University of Chicago ( <a href="http://www.norc.uchicago.edu">http://www.norc.uchicago.edu</a> )	Dan Black, Bob Michael, Kyle Fennel, Vicki Wilmer, Chuck Safiran, Mikhael Shoblum, Rupa Datta (NORC); Lisa Calderwood, Randy Banks, Geoff Angel
7-8 Apr 2009	NCHS/Westat, Hyattsville, MD ( <a href="http://www.cdc.gov/nchs/nhanes.htm">http://www.cdc.gov/nchs/nhanes.htm</a> and <a href="http://www.westat.com">http://www.westat.com</a> )	Lew Berman, Kathryn Porter, Debra Reed-Gillette, Jerry Del Rosso, George Zipf, Denise Schaar, Yechiam Ostchega, Rosemary Hirsch, Sam Notzon (NCHS), Taylor Cooper, Peggy Bowen, Debbi Hillard, Steve Bernas (Westat), Lisa Calderwood, Randy Banks, Geoff Angel
9 Apr 2009	NCES, Washington, DC ( <a href="http://nces.ed.gov/ECLS">http://nces.ed.gov/ECLS</a> )	Gail Muligan, Chris Chapman, Jill Barlibati (NCES), Lisa Calderwood, Randy Banks, Geoff Angel
10 Jun 2009	NatCen, London ( <a href="http://www.natcen.ac.uk">http://www.natcen.ac.uk</a> )	Carli Lessof, Reg Gatenby, Richard Boreham (NatCen); Lisa Calderwood, Jon Johnson
7 Oct 2009	German Institute for Economic Research, Berlin ( <a href="http://www.diw.de/en/soep">http://www.diw.de/en/soep</a> )	Juergen Schupp, Peter Krause, Jan Goebel, Ingo Sieber, Joachim R. Frick, Lisa Calderwood, Randy Banks
9 Oct 2009	GfK NOP, London ( <a href="http://www.gfknop.com">http://www.gfknop.com</a> )	Nick Moon, Lisa Calderwood, Randy Banks
13 Oct 2009	Ipsos MORI, London ( <a href="http://www.ipsos-mori.com/">http://www.ipsos-mori.com/</a> )	Kathryn Gallop, Lisa Calderwood, Randy Banks
15 Oct 2009	CentERdata, Tilburg ( <a href="http://www.centerdata.nl/en">http://www.centerdata.nl/en</a> )	Marcel Das, Eric Balster, Maurice Martens, Alerk Amin, Arnaud Wijnant, Jon Johnson, Randy Banks
20 Oct 2009	TNS-BMRB, London ( <a href="http://www.tns-bmrb.co.uk/">http://www.tns-bmrb.co.uk/</a> )	Michelle Harrison, Bill Blyth, Bruce Hayward, Lisa Calderwood, Randy Banks
21 Oct 2009	NatCen, London ( <a href="http://www.natcen.ac.uk">http://www.natcen.ac.uk</a> )	Hayley Cheshire, Reg Gatenby, Sue Brooker, Steve Kelly, Carli Lessof, Lisa Calderwood, Randy Banks

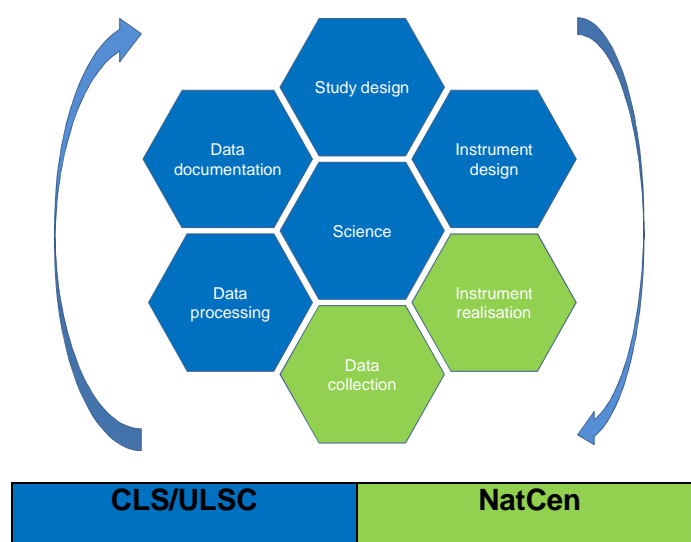
## Appendix E. Longitudinal surveys around the world<sup>78</sup>

This Appendix briefly summarises the major longitudinal surveys around the world from which information has been collected for this project, the ways in which the data production line is organised on these studies and, if applicable and known, the CAI software used on each study.

### 1958, 1970 and 2000-01 Cohort Studies and UK Household Longitudinal Study (UK)

The three cohort studies run by CLS and BHPS, and UKHLS run by ULSC are the major academic longitudinal studies in the UK and are funded primarily by ESRC. They are not described in detail in this report, as it is assumed that the intended readership will be familiar with them, but further information can be found at [www.cls.ioe.ac.uk](http://www.cls.ioe.ac.uk) and [www.iser.essex.ac.uk/ulsc](http://www.iser.essex.ac.uk/ulsc), respectively. The DPL is compartmentalised in the same way on all of these studies with CLS/ULSC mainly responsible for scientific direction, study design, instrument specification, data processing and data documentation; with instrument realisation and data collection the primary responsibility of a sub-contracted fieldwork agency. The fieldwork for these studies is competitively tendered periodically and NatCen is the incumbent fieldwork contractor on all of these studies<sup>79</sup>. The diagram below illustrates this division of responsibilities.

Diagram C1: The data production line on the UK cohort and panel studies



In practice, the compartmentalisation of the data production line is more 'fuzzy', e.g. the fieldwork agency may influence elements of the study design, comment on instrument specification, carry out some data processing and produce documentation of the data collection instrument; and CLS/ULSC are involved in the instrument realisation process through testing of the CAI instrument, overseeing the data collection process. As a result, there is some duplication of effort in the DPL on these studies, particularly at the interface between organisations, i.e. instrument design and realisation prior to data collection and

<sup>78</sup> NHANES was also reviewed but it is intentionally not discussed in this Appendix as it is not a longitudinal study.

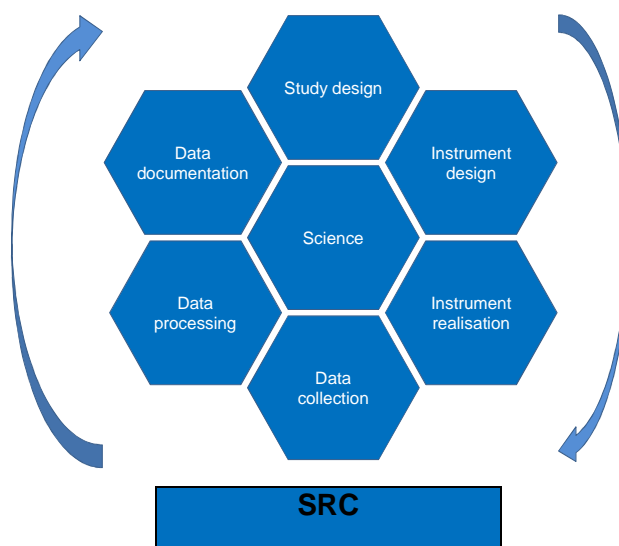
<sup>79</sup> NatCen have conducted fieldwork for the cohort studies almost exclusively over the last 10 years as a result of being the preferred bidder in six out of seven competitive tenders over the period. The exception was the second sweep of the MCS, in which Gfk-NOP was the preferred bidder. NatCen won the contract to carry out the first two waves of the UKLHS. The fieldwork for its forerunner, BHPS, has been carried out exclusively by Gfk-NOP since its inception in 1991.

data processing and documentation after data collection. The CAI software used on all of these studies is Blaise, as this is NatCen's preferred software.

### **Panel Study of Income Dynamics and Health and Retirement Study (USA)**

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal household panel study following almost 9,000 families. It is funded primarily by the National Science Foundation, and the data collected focus on the dynamics of economic and demographic behaviour. The Health and Retirement Study (HRS), begun in 1992, is a longitudinal panel study of people aged 50 and over, following around 22,000 individuals. It is funded primarily by the National Institute on Aging and the data collected focus on the interaction of health and economic circumstances in older age. For both studies, data is collected biennially using computer-assisted telephone interviewing (CATI). Further information can be found at <http://psidonline.isr.umich.edu/> and <http://hrsonline.isr.umich.edu/>. These studies are both run by the Survey Research Centre (SRC), Institute for Social Research, University of Michigan. All of the elements of the data production line are carried out within the context of a single institution.

**Diagram C2: The data production line on PSID and HRS**



There is an internal division within the organisation with the survey research operations (SRO) division essentially acting as in-house fieldwork agency for the PI team. The relative responsibilities of the PI team and the SRO were slightly 'fuzzy' and differed slightly between the two studies. For example, HRS has a CAI programmer within the scientific team (whereas this service is usually provided by SRO) and PSID have a Survey Director who's purview spans both parts of the organisation.

In this organisational model, the distance between the scientists and the other elements of the data production cycle is as short as possible. The primacy of the science is a key part of the organisational philosophy of the SRC, as is the importance of the survey method as a tool of scientific measurement and investment in interviewers as units of data production. A single organisation employs both the scientists for whom the data is collected and those collecting the data. The funding model under which these studies operate is that a single institution (SRC) raises the funds to carry out all elements of the data production line, i.e. the data collection is carried out internally rather than being competitively tendered by the PI team. As a result, there is also a long-term relationship between those collecting the data and those for and from whom the data is collected. In terms of efficiency, all elements of the data production being carried out by the same organisation is necessarily more efficient as

there is less duplication of work. The CAI software used on all of these studies is Blaise as this is SRC's preferred software.

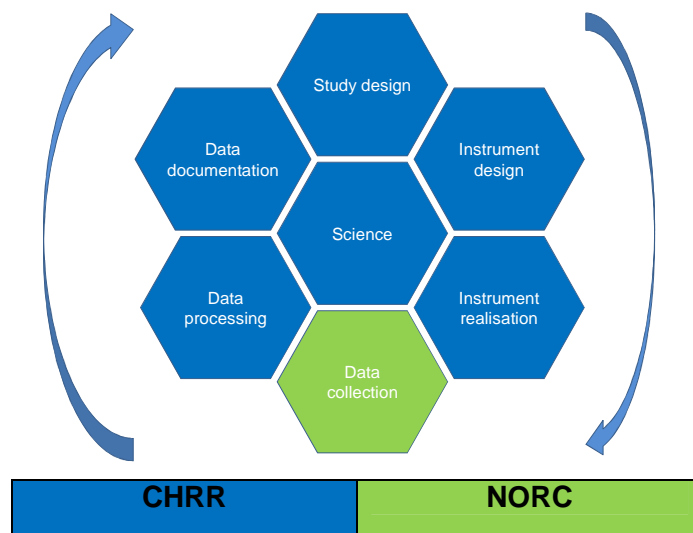
### The National Longitudinal Surveys (USA)

The National Longitudinal Surveys, funded by the US Bureau of Labor Statistics, track labour market activities over time. There are five studies, but the two core studies are those begun in 1979 and 1997. The National Longitudinal Survey of Youth 1997 (NLS97) follows a cohort of people born between 1980 and 1984 who were aged between 12 and 17 when first interviewed in 1997. The National Longitudinal Survey of Youth 1979 (NLS79) follows a cohort of people born between 1957 and 1964 who were aged between 12 and 22 when first interviewed in 1979. The NLS79 employs a sequential mixed mode design with a telephone phase followed by a face-to-face phase on a biennial basis. The NLS97 is a face-to-face study conducted annually. More information can be found at <http://www.bls.gov/nls/>. These studies are carried out jointly by CHRR at Ohio State University and NORC at the University of Chicago. On both the NLS79 and NLS97, CHRR are primarily responsible for instrument realisation, data processing and data documentation with NORC responsible for the data collection. On the NLS79, CHRR are also primarily responsible for the scientific direction, study design and instrument design; whereas for the NLS97, these areas are the joint responsibility of both organisations.

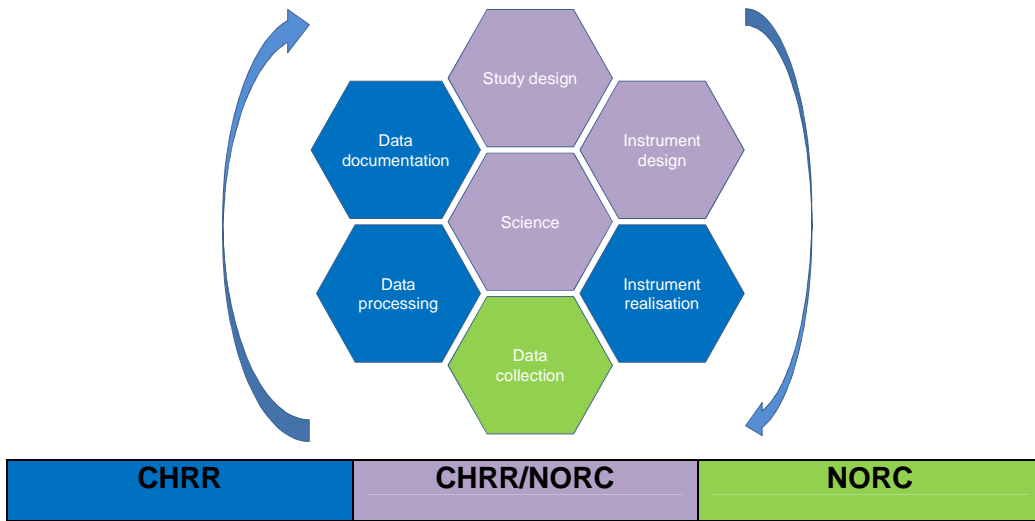
The relationship between CHRR and NORC on these studies is close and long-standing. Both organisations jointly bid to carry out the study and rotate periodically as the 'prime' contractor. The fieldwork is not competitively tendered. Rather, a fieldwork agency is a joint and equal partner in the study. This enduring collaboration means that both organisations are highly invested in the scientific aims of the studies and have a shared focus on the end-product, i.e. the data.

This division of the DPL means that the same organisation (CHRR) is responsible for instrument realisation as well as data processing and documentation. In consequence, they have invested significantly over a long period of time in bespoke software (SurveySuite) designed to closely integrate these elements of the DPL. Bringing the CAI instrument and the documentation of the data produced by it closer together has undoubtedly brought many benefits to these studies. However, this particular division of the DPL is unusual and developed as a result of circumstances idiosyncratic to these studies. In addition, the software used to achieve this is bespoke to these studies and is not used outside the organisation which developed it. In general, it is not common for the data collection instrument to be produced by a different organisation from the organisation responsible for carrying out the data collection.

**Diagram C3: The data production line on NLS79**



**Diagram C4: The data production line on NLS97**

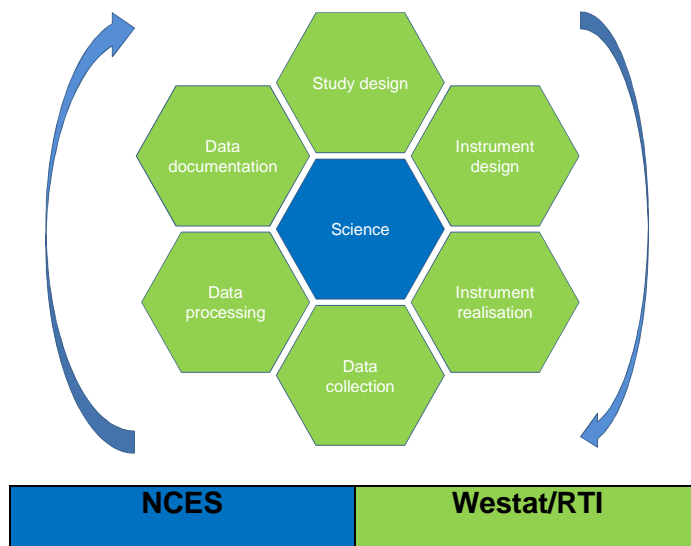


**The Early Childhood Longitudinal Study (USA)**

The Early Childhood Longitudinal Study (ECLS) programme includes three longitudinal studies that examine child development, school readiness and early school experiences. The birth cohort (ECLS-B) is a sample of children born in 2001 and followed from birth through kindergarten entry. The kindergarten class of 1998-99 (ECLS-K) cohort is a sample of children followed from kindergarten through the eighth grade. The kindergarten class of 2010-11 (ECLS-K:2011) cohort will follow a sample of children from kindergarten through the fifth grade. Further information can be found at <http://nces.ed.gov/ECLS/>.

This study is managed by the National Center for Education Statistics (NCES). Although they retain overall responsibility for the scientific direction of the study and contribute substantially to the study design process, all of the other elements of the DPL are sub-contracted to the fieldwork agencies, currently Westat (ECLS-K) and RTI (ECLS-B). NORC also played a significant role in the study design of ECLS-K. On occasion, they also sub-contract research to inform the scientific direction of the study to Westat/RTI, and work very closely with the Education Statistics Services Institute on most aspects of the process supporting the scientific core.

**Diagram C5: The data production line on ECLS**

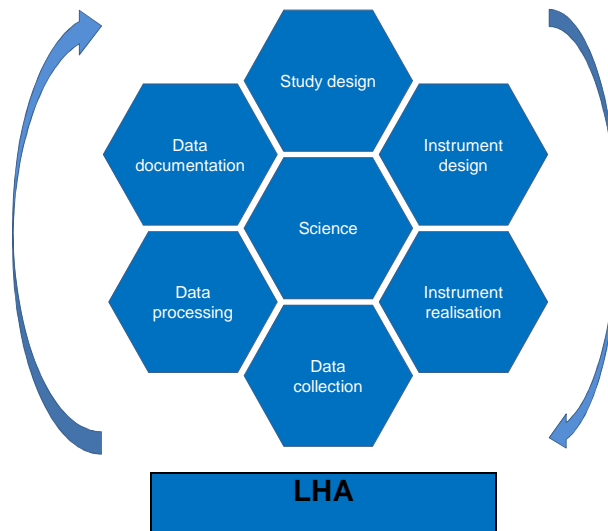


In this model almost all of the elements of the data production line are carried out by the same organisation – in this case, it is a fieldwork agency. However, they have very little autonomy and are closely monitored and controlled by NCES through carefully specified and detailed contracts. The CAI software used on these studies is Blaise, one of several CAI packages supported by Westat/RTI.

### **The National Survey of Health and Development (UK)**

The National Survey of Health and Development (1946 cohort) is the oldest and longest running of the British birth cohort studies. From an initial maternity survey of 13,687 (82%) of all births recorded in England, Scotland and Wales during one week in 1946, a socially stratified sample of 5,362 singleton babies born to married parents was selected for follow-up. This sample comprises the NSHD cohort and participants have been studied 21 times. The study is run by the MRC Unit for Lifelong Health and Ageing (LHA). All aspects of the DPL are carried out by this team. The data collection on this study is not carried out using CAI methods.

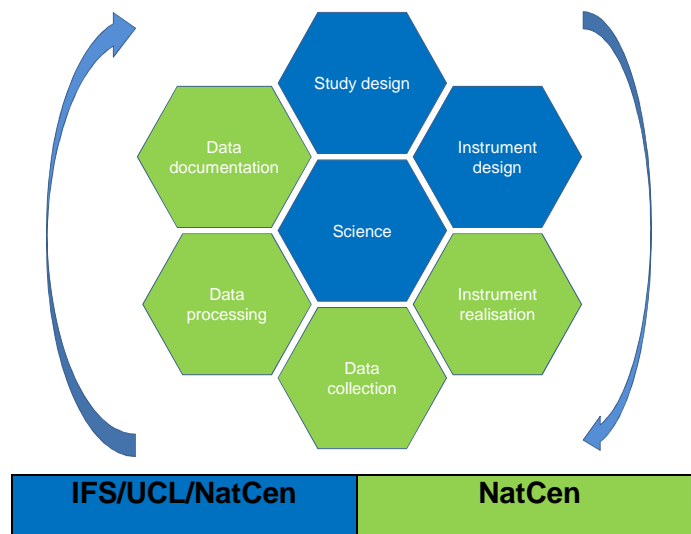
**Diagram C6: The data production line on the 1946 cohort**



### **English Longitudinal Study of Ageing (UK)**

The English Longitudinal Study of Ageing (ELSA) is a panel study of over 12,000 people aged 50 and over which started in 2002. It is carried out jointly by the Institute for Fiscal Studies (IFS), Department of Epidemiology and Public Health at University College London (UCL) and NatCen. The CAI software used is Blaise.

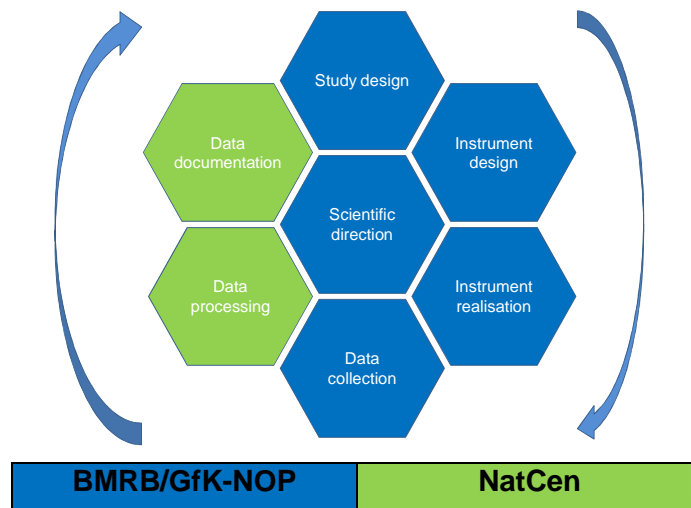
**Diagram C7: The data production line on ELSA**



**Longitudinal Survey of Young People in England (UK)**

The Longitudinal Survey of Young People in England (LSYPE) follows over 15,000 children who were in Year 9 (aged 13/14) in 2004. It is funded by the Department of Children, Schools and Families. Several elements of the data production line are sub-contracted to a consortium of fieldwork agencies led by British Market Research Bureau (BMRB), which also includes GfK-NOP. Different elements of the DPL are carried out by NatCen. SPSS Dimensions is the CAI software used on this study.

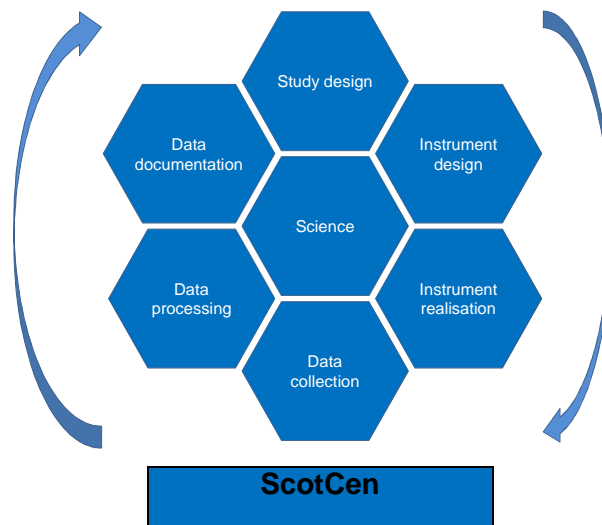
**Diagram C8: The data production line on LSYPE**



**Growing Up in Scotland (UK)**

Growing Up in Scotland (GUS) follows the lives of 8,000 children in Scotland annually. It consists of two cohorts: one of 5,000 children, born between June 2004 and May 2005, who were recruited as babies (around 10 months); and one of 3,000 children, born between June 2002 and May 2003, who were recruited as toddlers (around 34 months). The study is run by the Scottish Centre for Social Research (ScotCen), which is responsible for all aspects of the study, in collaboration with Centre for Research on Families and Relationships, at the University of Edinburgh, and the MRC Social and Public Health Sciences Unit at Glasgow University, who primarily provide input into the scientific direction. Data collection is carried out primarily by CAPI and the software used in Blaise.

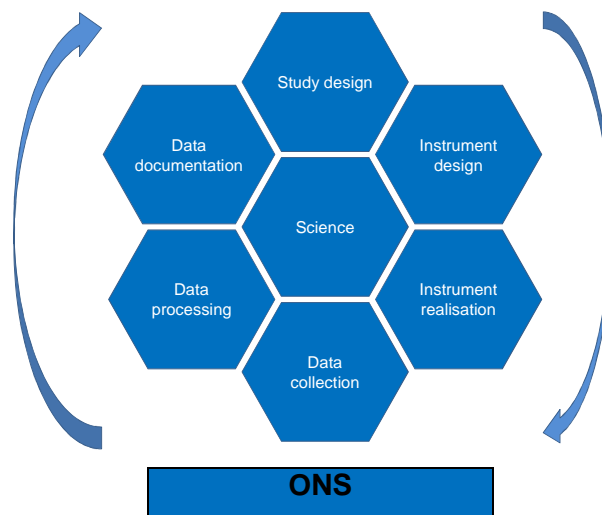
**Diagram C9: The data production line on GUS**



**Wealth and Assets Survey (UK)**

The Wealth and Assets Survey (WAS) collected information about the economic well-being of households and individuals and has a target sample size of 32,000 adults in Great Britain. All aspects of the study are carried out by the Office for National Statistics (ONS) with input from the other funding organisations –Department for Work and Pensions; Department for Business, Enterprise and Regulatory Reform; HM Treasury; HM Revenue & Customs; Department for Communities and Local Government; and the Cabinet Office. The CAI software used is Blaise.

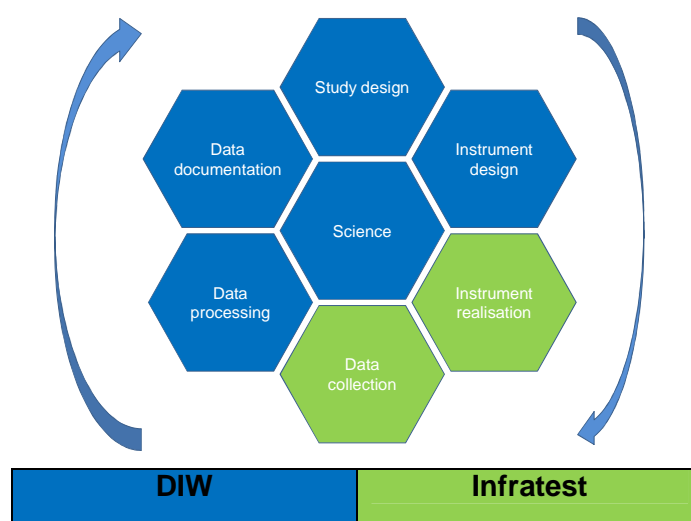
**Diagram C10: The data production line on WAS**



**German Socio-Economic Panel (Germany)**

The SOEP is an annual panel study of around 20,000 people in 11,000 private households. It is run by the German Institute for Economic Research, DIW Berlin. Mixed mode data collection is carried out by TNS Infratest Sozialforschung. The CAI software used is Intuitive.

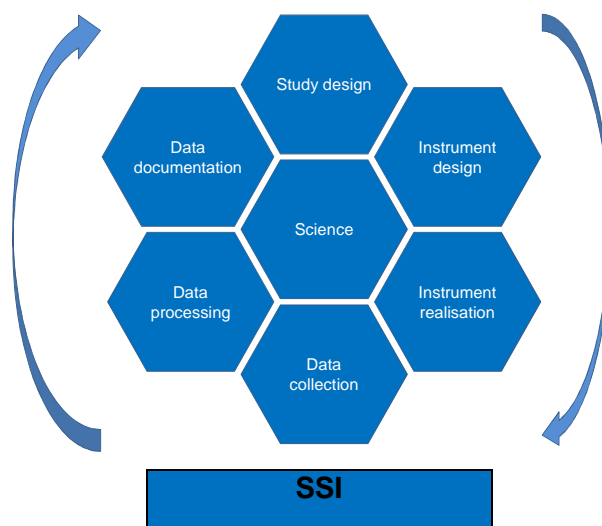
**Diagram C11: The data production line on SOEP**



**Danish National Birth Cohort (Denmark)**

The Danish National Birth Cohort (DNBC) recruited around 100,000 pregnant women between 1997 and 2002. The study is run by the Danish Epidemiology Science Centre at Statens Serum Institut (SSI) in Copenhagen. All aspects of the DPL are carried out by this team. The data collection on this study is carried out by CATI and postal methods. More information can be found at <http://www.ssi.dk/sw9314.asp>.

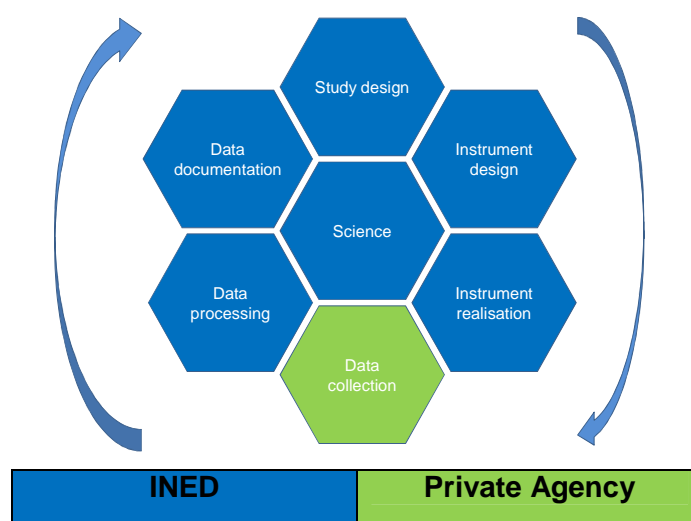
**Diagram C12: The data production line on DNBC**



**Growing Up in France (France)**

Growing Up in France or Étude Longitudinal Française depuis L'Enfance (ELFE) is aiming to recruit 20,000 children born in France in 2009. The study is run by the Institut National D'Études Démographiques (INED). CATI data collection will be sub-contracted to a private agency. In the pilot study, the CAI software used was VOXCO.

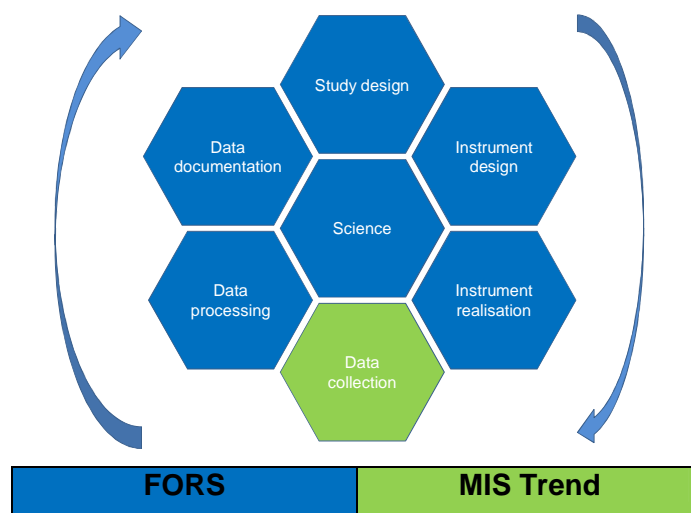
**Diagram C13: The data production line on ELFE**



**Swiss Household Panel Study (Switzerland)**

The Swiss Household Panel (SHP) is an annual panel study following a random sample of households in Switzerland over time, interviewing all household members. Data collection started in 1999 with a sample of 5,074 households containing 12,931 household members. In 2004 a second sample of 2,538 households with a total of 6,569 household members was added. The study is run by the Swiss Foundation for Research in Social Sciences. Data collection is carried out by CATI and sub-contracted to M.I.S. Trend.

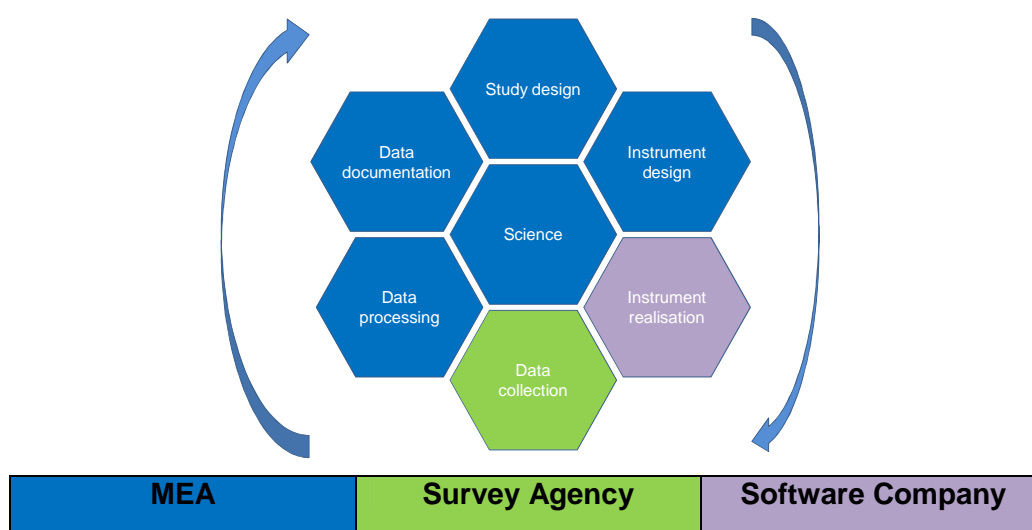
**Diagram C14: The data production line on SHP**



**Survey of Health, Aging and Retirement in Europe (Europe)**

The Survey of Health, Ageing and Retirement in Europe (SHARE) is a multidisciplinary and cross-national panel survey of more than 45,000 individuals in 16 European countries aged 50 or over. SHARE is co-ordinated by the Mannheim Research Institute for the Economics of Aging (MEA). The CAI software used is Blaise. Data collection is sub-contracted to commercial survey agencies in each country and instrument realisation is sub-contracted to a software company.

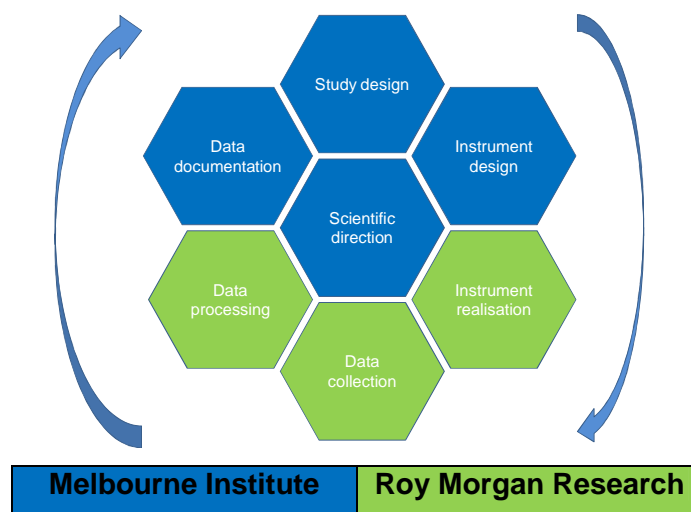
**Diagram C15: The data production line on SHARE**



**Household, Income and Labour Dynamics in Australia (Australia)**

The Household, Income and Labour Dynamics in Australia (HILDA) Survey is a household-based panel study which began in 2001. The wave 1 panel consisted of 7,682 households and 19,914 individuals. Interviews are conducted annually with all adult members of each household. The HILDA Survey was initiated, and is funded, by the Australian Government through the Department of Families, Housing, Community Services and Indigenous Affairs. Responsibility for the design and management of the survey rests with the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne. Data collection for waves 1 to 8 was sub-contracted to The Nielsen Company, a private market research company. Data collection for waves 9 to 12 will be undertaken by Roy Morgan Research. Conformat is the CAI software used on this study.

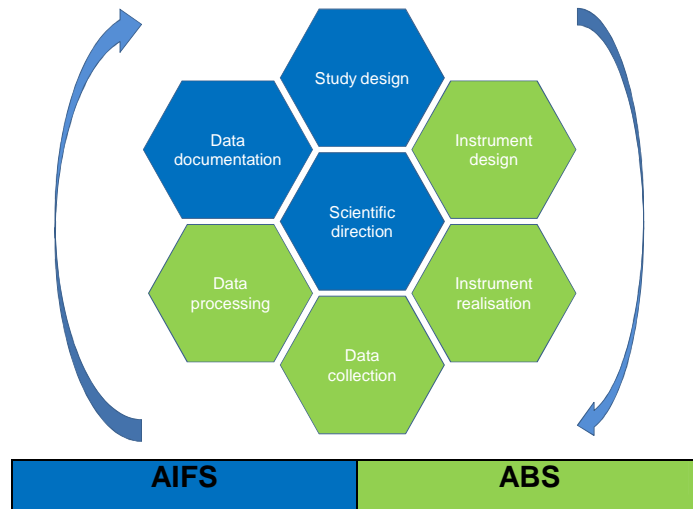
**Diagram C16: The data production line on HILDA**



**Growing Up in Australia, the Longitudinal Study of Australian Children (Australia)**

The Longitudinal Study of Australian Children (LSAC) collects data every two years from two cohorts of 5000 children. The first cohort was aged 0-1 years in 2003/4 and the second cohort was aged 4-5 years in 2003/4. The study is funded by the Australian Government Department of Families, Housing, Community Services and Indigenous Affairs and run by the Australian Institute for Family Studies (AIFS) in Melbourne. Data collection is carried out primarily by CAPI by Australian Bureau of Statistics (ABS). Blaise is the CAI software used on this study.

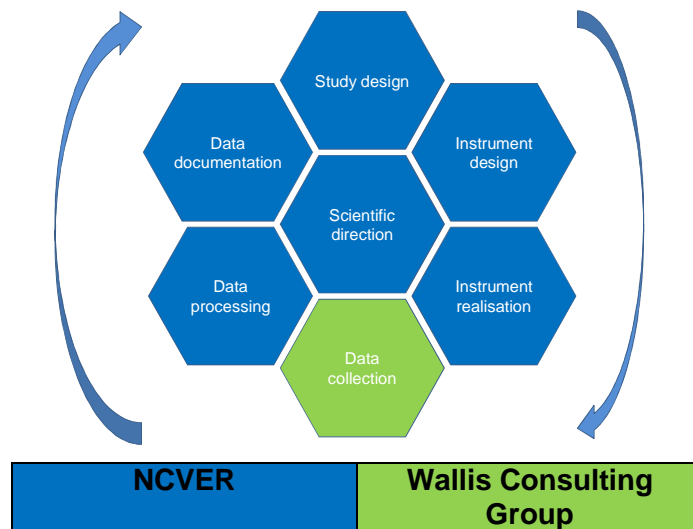
**Diagram C17: The data production line on LSAC**



**Longitudinal Surveys of Australian Youth (Australia)**

The Longitudinal Surveys of Australian Youth (LSAY) are a series of cohort studies, begun in 1995, each following around 10,000 children from age 15 for 10 years. The study is funded by the Department of Education, Employment and Workplace Relations and run by the National Centre for Vocational Education Research (NCVER) at the University of Adelaide. The data collection is carried out by CATI by the Wallis Consulting Group. Quantum is the CAI software used on this study.

**Diagram C18: The data production line on LSAY**



## Appendix F. International Survey of Major Longitudinal Studies

### Division of responsibilities and relationship with other organisations

1. On the most recent wave of your study, were any of the following tasks carried out mainly by an organisation other than your own: (*Tick all that apply*):

*Scientific Direction, Study Design, Instrument Design, Instrument Realisation, Data collection, Data processing, Data documentation, Sample maintenance (between waves)*

2. What is/are the name(s) of the organisation(s)? (*max 3*)

### Organisation of work and relationship with other organisations

3. (On the most recent wave of your study,) which of the following tasks were mainly carried out by Organisation 2.a)? (*Tick all that apply.*)

*Scientific Direction, Study Design, Instrument Design, Instrument Realisation, Data collection, Data processing, Data documentation, Sample maintenance (between waves)*

4. Which of these contractual arrangements best describes your relationship with Organisation 2.a) (on the most recent wave of your study)?

*They were sub-contacted to us, We were sub-contacted to them, We were equal partners, Other (specify)*

5. Approximately how many years and/or months does your current (or most recent) contract with Organisation 2.a) cover?

6. How many waves of data collection does your current (or most recent) contract with Organisation 2.a) cover?

7. How many organisations submitted tenders for this sub-contract (when it was last put out to tender)

8. Was your current (or most recent) sub-contract with Organisation 2.a) subject to competitive tender? (*Yes/No*)

9. How many organisations submitted tenders for this sub-contract (when it was last put out to tender)?

10. Which of the following cost-reimbursement models is used in your current (or most recent) sub-contract with Organisation 2.a)?

*Fixed cost, Fixed cost plus variable fees, Other (specify)*

11. How detailed would you say your current (or most recent) sub-contract with Organisation 2.a) is in relation to the responsibilities of each organisation?

*Very detailed, Quite detailed, Not very detailed, Not at all detailed*

### CAI Software and instrument development

12. Was any of the data collection for the most recent wave of your study carried out using Computer Assisted Interviewing (CAI)? (*Yes/No*)

13. Which CAI software was used?

14. What was the main method by which CAI instruments were specified to the programmer?

*Microsoft Word document, Specifically designed CAI specification tool (specify), Other method (specify)*

15. Which of the following methods were used to test the CAI instruments? Please tick all that apply.

*Manual checking of CAI instrument against specification, Data checking including data flooding, Scenario testing, Other (specify)*

### Relational Databases

16. Is any Relational Database Management Software (RDBMS) used on your study? (*Yes/No*)

17. Which of the following is RDBMS software used for? Please tick all that apply.

*Management and processing of survey data, Management and processing of sample information, Other (specify)*

18. Which RDBMS software is used for management and processing of survey data use as mentioned in the previous question?
19. Which RDBMS software is used for management and processing of sample information use as mentioned in the previous question?
20. Which RDBMS software is used for 'other' use as mentioned in the previous question?

### **Interviewers and fieldwork**

21. On the most recent wave of your study, did the interviewers need to pass a formal test following their training to demonstrate their understanding of the procedures before commencing work on the study? (Yes/No)
22. On the most recent wave of your study, was the data from the first x interviews from each interviewer checked on a real-time basis and problems fed back the interviewer? (Yes/No)
23. On the most recent wave of your study, did your organisation have access to real-time information about fieldwork progress? (Yes/No)
24. What level was this information available at? *Please tick all that apply.*

*Case-level, Interviewer-level, Area-level, Other aggregate level*

### **Data and documentation**

25. On the most recent wave of your study, was the data delivered to you in real-time? (Yes/No)
26. On the most recent wave of your study, which, if any, of the following kinds of paradata (data about the survey process) were delivered to you? *Please tick all that apply.*

*Case-level outcome codes, Individual records of interviewer contact attempts, Individual records of interviewer tracing attempts, Details of each time the case was issued (for cases issued multiple times), (Anonymised) interviewer ID, Demographics of individual interviewers e.g. age, ethnicity, Measure of experience and ability for individual interviewers e.g. length of service, grade, Data collected from quality control procedures, Other (specify), None of these*

27. How are the data and documentation from your study made available to external end users? *Please tick all that apply.*

*From a national data archive, From one of the organisations involved in running the study, Other (specify), Not applicable: Data is not available to external end users*

28. What, if any, standards are supported in the production of documentation e.g. DDI?

### **User Support**

29. What kinds of support and training are provided for end users? Please tick all that apply.

*Classroom-based training workshops and/or courses, Individual tutorials which can be followed by users in their own time, One-to-one support e.g. in response to specific queries, Other (specify)*