Young Lives Methods Guide

Data Management: Cleaning, Storage and Archiving

July 2011

Young Lives aims to produce high quality data about the lives of children living in poverty for long-term use. Strong research is built on careful and consistent data management. This includes:

- designing programmes and databases for data entry
- preparing data for analysis by careful inputting and cleaning
- ensuring that data are stored so that they can be easily accessed, analysed and interpreted by researchers
- preparing data for public archiving.

The longitudinal aspect of the Young Lives study results in multiple, linked rounds of fieldwork which present special challenges for data management. Changes to the survey questionnaires are made in each round as the children grow older. Before any round of fieldwork begins, data managers have to ensure that they have set up systems to manage the data through all stages, from tracking and collection to cleaning and archiving. These systems not only have to be consistent across countries, but also between different rounds of survey and qualitative research.

The sheer quantity of Young Lives data also makes data management challenging: a single one of the five planned rounds of survey data comprises the answers to more than 12,000 questionnaires, each with over 30 pages of questions, applied in 80 sites in 4 countries. In the first two rounds of the survey, all questionnaires were on paper. It usually took up to two hours to enter the data from one questionnaire in a database. In addition to the main survey, Young Lives also produces a mass of qualitative data in a variety of formats including field notes, audio recordings, interview transcripts, photographs and videos.

Storing this quantity of data securely and ensuring the confidentiality of respondents are particularly important aspects of data management in the Young Lives study. By 2007, the size of the qualitative and quantitative databases was so great that data managers needed to build a WebDAV server – accessed by a password-protected, encrypted website – for sharing files securely between study countries. By 2009, some field teams were piloting electronic data collection methods to increase the efficiency and security of data collection and management.

Processing and storing survey data

Each Young Lives country team has a full-time data manager who works closely with the Oxford-based data manager. For the first two rounds of the survey, each questionnaire was handled by a fieldworker, a supervisor and two data entry clerks. Their collective task was to implement systematic and consistent methods of data recording, labelling and storage, to ensure that data were clean, secure, backed up and confidential.

For each round of the survey, a database is built from the questionnaire using Microsoft Access, with country-specific sections and questions added by each team. Before fieldwork begins, country data managers run the database with test data to check its internal consistency and report on problems. Once the database is revised, data managers train data entry staff, a process which often generates more changes and refinements. The database and questionnaire are finalised before the survey begins, and a copy of the database is saved onto each data entry computer.

When the survey team has produced, completed and checked the questionnaires, data are entered onto the computers and then merged into a single database. The same data are then re-entered by another person, producing a second database. This is compared to the first database using EpiInfo software to check for entry errors. Data managers check all queries against the hard copy of the questionnaire and both databases are updated with corrections.

Subsequent stages of data cleaning focus on highlighting inconsistencies. Data managers flag inconsistencies in the databases and then either go back to fieldwork teams or check against hard copies to find out why they have arisen and fix them. Later, as researchers begin their analysis of the data, other contradictions and unusual patterns emerge that also need to be checked. The bulk of data-cleaning work is concentrated over a few months, during which time the data is all transferred to SPSS, the software in which it is publicly archived. A cut-off date for data cleaning is decided according to the date set for public archiving.

Systems and processes for managing and storing survey data have been developed and refined through each survey round as new issues and challenges have emerged:

- In Round 1 of the survey (2002), all teams found errors in initial databases that indicated inadequate checking prior to data-entry, so more training and time were given to this in subsequent rounds. Data-checking for inconsistencies in SPSS led to some errors in the Access database having to being corrected, so more attention was given in later rounds to running error trapping programmes in the database before transfer to SPSS.

- In Round 2 (2006–07), the databases had to be built off a new and larger questionnaire which went through multiple revisions. Minor changes made to the questionnaires after the databases were built sometimes meant major, last-minute programming changes. Double entry of all survey data was adopted for increased accuracy. Consistency checks were run in Access and the data cleaned before transfer to SPSS. Round 2 data were used to further cross-check Round 1 demographic information.
- In Round 3 (2009), Young Lives began to shift from paper to electronic data collection. Personal Digital Assistants (PDAs) were used to collect over 70% of the data in Vietnam and 50% in Peru. The PDAs used bespoke software built for dealing with a complex survey. This software was piloted alongside the Round 3 questionnaire, which included a new sibling component and a self-administered section for the older cohort of the sample. The remaining surveys in Vietnam and Peru, as well as all surveys in India and Ethiopia, used paper based questionnaires in this round, and followed the same process as Round 2.
- In Round 4 (2013), Young Lives aims to collect data electronically in all four countries. Some of the work of building software and databases will be a joint effort of the central data managers and country teams.

Processing and storing qualitative data

In addition to the survey, Young Lives has generated qualitative data using interviews, observation and group methods, all of which are audio recorded. The resulting dataset is both multimedia and multilingual. Most qualitative data are handled by one or more fieldworkers, a transcriber and a translator. Preparing this data for storage has required teams to agree on and implement systematic methods for data recording, setting up documents and labelling data files in all formats, and a consistent file and folder structure within and across countries. Following security and confidentiality protocols is also essential. While qualitative data are not archived publicly, they nonetheless must be prepared for external scrutiny and audit as part of the process of the study.

A set of guidelines for managing qualitative data across all research teams was established in 2007. It sets out a standardised naming structure for each file which facilitates quick identification of necessary information, a standardised format for text documents, and a Record Header detailing the circumstances of the interview which has to be attached to all data in all formats. It also includes objectives and protocols for transcription and translation of audio recordings.

Qualitative data are stored in a range of file formats before being analysed using Atlas.ti, software that was selected because it can display and process text, graphic, audio and video data, and keep track of notes, annotations, codes and memos. Researchers use the programme to build a coding structure for data analysis.

Qualitative data must be checked both for its consistency with quantitative data, and across successive qualitative rounds. The process of cleaning and checking qualitative data is made more complicated by the issue of accurate translation and transcription. Across the whole study, research is carried out in nearly thirty languages, so going back to check data against the original audio files takes considerable time and resources and involves several people. Sometimes certain phrases or sections of transcribed material are double-checked, but this is opportunistic rather than systematic.

Public archiving

There are ethical questions raised by deciding whether or not to deposit data in public archives. Protecting the identity of Young Lives children and their communities is a key principle of the study's approach to ethical research. While survey data can be anonymised to achieve this protection, this is much more difficult, time-consuming and costly to achieve with qualitative data. For these reasons, Young Lives only publicly archives its survey data.

Why share data in public archives?

- to allow other researchers to replicate, validate, correct and build on results, which strengthens scientific integrity
- to enable the exploration of themes not anticipated by the researchers who started the study
- to prevent duplication of research
- to provide wider visibility for research processes and their findings
- to increase the impact of research within and beyond the discipline, sector or country where it was carried out
- to preserve research for the long-term future
- to meet the requirements of many funding agencies and journals which now request data to be publicly archived to increase the returns to publicly funded research.

Young Lives survey data are archived with the Economic Social and Data Service (ESDS), the UK national data archiving and dissemination service. All data are anonymised and any text-based variables are deleted prior to archiving. For each of the four countries, the datasets comprise SPSS files of individual and household information for each of the two cohorts of children, including key composite variables such as a household wealth index and mental health scores. There are also several sub-files which include household rosters and data at a lower level.

To contextualise these datasets, Young Lives has deposited a range of other information and documentation in the archive:

- A short background to the study, which explains the structure of the survey and the main topics of different rounds
- Household, child and community questionnaires for each round, together with justification documents describing what questions were asked and how they were arrived at
- An explanation of sampling procedures
- Fieldworker manuals
- A data dictionary that describes each variable
- The method of calculation for calculated variables
- Consent forms for each round.

Before accessing the data, users must register and apply for a password with ESDS and sign a confidentiality agreement. They are also asked to inform ESDS and Young Lives of any analysis or publications resulting from their work with the dataset.

Challenges and lessons

Challenges faced by data management in storing and archiving Young Lives data include:

- Selecting the right software and hardware for different aspects of data management, and training all staff in their use
- Ensuring adequate time to build databases and enter questionnaires
- Coordinating the piloting and revision of questionnaires and databases
- Coordinating with other study team members to ensure that everything needed for fieldwork is ready in time to ensure that cohorts are interviewed when children are the right age
- Achieving consistency in complex data management processes across four culturally and linguistically different settings
- Achieving consistency between different rounds of the survey, and linking data between rounds
- Giving researchers rapid access to data to generate policy-relevant results without compromising the need to carry out validation checks on the data.

Team management is a vital aspect of meeting these challenges – both within the data management team, and between data managers and other members of the wider study team. Oxford-based data managers have increasingly tried to create opportunities for country team members to share concerns, ideas and ways of working from country to country, instead of just through bilateral communications with Oxford. Country data managers report on their activities more systematically than before, through a data management section in each quarterly country report. Secure transfer of data between data managers and researchers has now been formalised, and data management is on the agenda of all team meetings.

Further reading

The UK Data Archive website (http://www.data-archive.ac.uk) provides information and advice on documenting, formatting and storing data. Supporting documentation for the Young Lives study can be found at http://www.esds.ac.uk/findingData/snDescription.asp?sn=5307#doc.