Harmonizing the data collection and data entry applications for longitudinal and cross-sectional surveys in social science: A metadata driven approach

Benjamin D Clark and Gayatri Singh

Aim of the paper

This paper emphasizes the necessity to take into account the *life cycle of data* throughout the various stages of a survey based research project. Applying this to the case of longitudinal and cross-sectional survey research, we provide a framework for thinking of data documentation as an integral component of the whole process that facilitates each subsequent stage of a research design, data collection, data capture and extraction. After laying out the data life cycle approach as the conceptual basis of thinking about survey design, this paper discusses two of its aspects in greater detail, namely, the *data collection instrument* (conceptual development and physical design), and the *data entry aspects.* Finally, we conclude the paper with suggestions towards and an exposition of a metadata approach to facilitate avenues for data sharing via better data documentation procedures, a goal that remains elusive to much of the data collection in the social science realm, particularly research that involves repeated cycles of data collection.

Exposition of the stages in the Life Cycle of Data:

Following seven stages should be seen as an integral part of running a robust study, especially with respect to longitudinal data collection systems. Due to space constraints we have only provided the stages in the bulleted and visual form. A comprehensive exposition will be included in the full paper.

Stage 1: Communication with the data system of the project site

- Researchers familiarize themselves with data system
- Clearly set out the forthcoming demands on the data system
- Set up efficient modes of communication and chains of authority
- Enumerate adequate resource allocation for data management

Stage 2: Data collection instrument (DCI) initialization and development

- Metadata driven instrument design with a generalized format
- Customized to project needs (paper, PDA, WWW etc)

- Accompanying data dictionary automatically generated
- Automated quality and consistency checks built in
- Form Flow and management of instrument established

Stage 3: DCI Generation and Flow Management

- DCI inventory management system applied
- Each DCI registered and tracked from initialization to archiving
- Tasks include: barcodes, other identifiers, sampling, archiving, merging
- Widespread adaptability to instrument types

Stage 4: Data collection

- Fielding of the instrument, data entry and archiving
- Metadata based data entry, quality checks, progress reports
- Final data preparation for the operational database and archives

Stage 5: Archiving

- Essential and efficient storage of source data, metadata and captured data
- Pr-planned to allow for inevitable complex corrections in the future

Stage 6: Analytical dataset generation

- Generated from the data archive repository
- A collection of standardized processes used
- Metadata and the captured data utilized
- Customized analytical datasets with speed and accuracy

Stage 7: Project output production

- Outputs and publications tracked
- Done with respect to each analytical dataset, by each project and researcher
- Key step for monitoring and demonstrating productivity

Please see Figure 1 (at the end) for a more intuitive visual description

Illustrative discussion of two aspects of the data life cycle

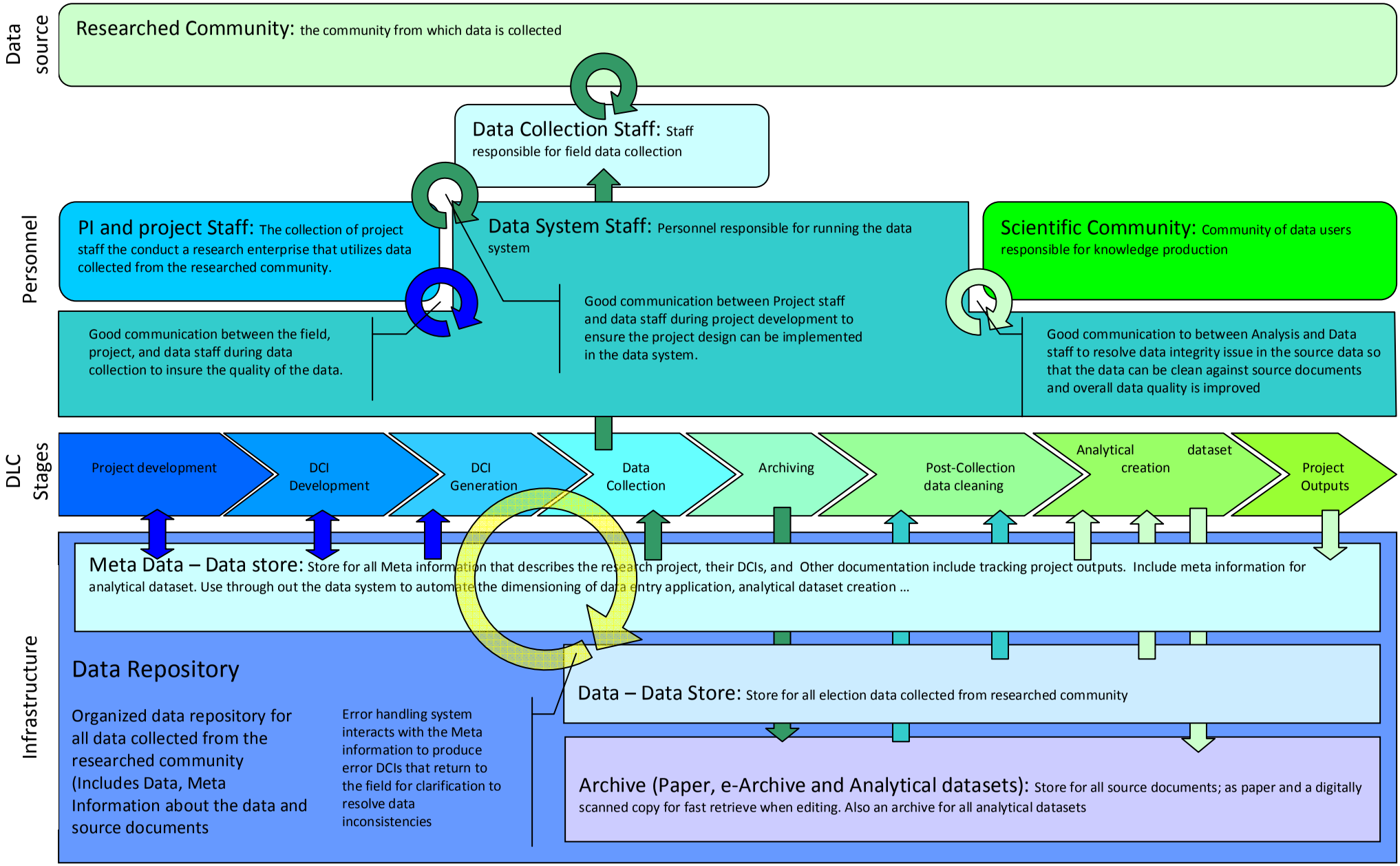
Textbook treatment of these two stages, namely data collection stage and the data entry stage, is not often closely linked. In fact, the data entry aspects are at times wholly missing from the readings on survey methodology, almost seen as an inevitable and a necessary step in the culmination of the research process but not as something that should be harmonized with the previous stages in the data life cycle. We stress a questionnaire layout that is harmonized with the data entry application, such that the latter is embedded within the same logical framework as the former. We also suggest tools and templates for the structural development of a questionnaire that further aids this kind of harmonization. Related to this, we provide insights into questions regarding the development of *individual identification systems* linked to the data collection instruments, particularly important in the case of longitudinal data collection following the same individuals over time.

Metadata approach and data documentation

Finally, we discuss a simple metadata driven approach for maintaining data documentation throughout the survey process that would enable effective data sharing. Such documentation issues should be especially important to social science research where the potential analysts of the collected data may not be pre-identified or even come from the same disciplinary background as that of the primary data collectors. The innovation of this approach lies in harmonizing the design of the questionnaire with the data entry software that utilizes a SQLExpress database and a meta-data driven, self documenting front end written in VB.Net. These suggestions are particularly important for low budget projects, especially those using multiple language translations as significant cost can be reduced in terms of data typist time, data cleaning activities and generating documentation for datasets. So far, this approach incorporating a metadata driven data collection and harmonized data entry application within a holistic and integrated data documentation framework has been tested in its various stages in a cross sectional survey in urban South Africa as well in a demographic surveillance site in rural South Africa collecting longitudinal data. As with any new approach, many lessons were learnt during the application of this method. Most acutely felt was the need to develop a better form flow management system and problematic questionnaire error handling. The lessons learnt and the subsequent improvements made will be shared as insights in final paper.

A metadata approach

Finally, we discuss a simple metadata driven approach for maintaining data documentation throughout the survey process that would enable effective data sharing. Such documentation issues should be especially important to social science research where the potential analysts of the collected data may not be pre-identified or even come from the same disciplinary background as that of the primary data collectors. The innovation of this approach lies in harmonizing the design of the questionnaire with the data entry software that utilizes a SQLExpress database and a meta-data driven, self documenting front end written in VB.Net. These suggestions are particularly important for low budget projects, especially those using multiple language translations as significant cost can be reduced in terms of data typist time, data cleaning activities and generating documentation for datasets. As with any new approach, many lessons were learnt during the application of this method. Most acutely felt was the need to develop a better form flow management system and problematic questionnaire error handling, which is now being investigated by the first author as an improvement within the data documentation process.


Data source
Researched Community: the community from which data is collected
Data Collection Staff: Staff responsible for field data collection
Personnel
PI and project Staff: The collection of project staff the conduct a research enterprise that utilizes data collected from the researched community.
Data System Staff: Personnel responsible for running the data system
Scientific Community: Community of data users responsible for knowledge production
Good communication between the field, project, and data staff during data collection to insure the quality of the data.
Good communication between Project staff and data staff during project development to ensure the project design can be implemented in the data system.
Good communication to between Analysis and Data staff to resolve data integrity issue in the source data so that the data can be clean against source documents and overall data quality is improved
DLC Stages
Project development
DCI Development
DCI Generation
Data Collection
Archiving
Post-Collection data cleaning
Analytical dataset creation
Project Outputs
Infrastructure
Meta Data – Data store: Store for all Meta information that describes the research project, their DCIs, and Other documentation include tracking project outputs. Include meta information for analytical dataset. Use through out the data system to automate the dimensioning of data entry application, analytical dataset creation …
Data Repository
Organized data repository for all data collected from the researched community (Includes Data, Meta Information about the data and source documents
Error handling system interacts with the Meta information to produce error DCIs that return to the field for clarification to resolve data inconsistencies
Data – Data Store: Store for all election data collected from researched community
Archive (Paper, e-Archive and Analytical datasets): Store for all source documents; as paper and a digitally scanned copy for fast retrieve when editing. Also an archive for all analytical datasets