Managing the Kanchanaburi Demographic Surveillance System: Creation of a Relational Database Management System

Our experience indicates that it is essential when creating any longitudinal database to invest in the development of systems that maintain confidentiality, while affording the basis for the numerous data linkages that are required for longitudinal data analysis.

By Jongjit Rittirong*

Increasing attention is being paid to the collection of longitudinal data. This attention is, in part, a response to the difficulties faced in establishing causal relations with cross-sectional data. However, the collection and use of longitudinal data has brought with it a series of challenges that are not faced by researchers manipulating cross-sectional data. In this article we describe how initial data management models, based on cross-sectional data storage and

* Institute for Population and Social Research, Mahidol Universtity, Salaya, Nakhon Pathom, Thailand, e-mail: prjrt@mahidol.ac.th.

manipulation used in the Kanchanaburi Demographic Surveillance System (KDSS), were found to be inadequate and were replaced by a database system that is consistent with longitudinal data collection, storage and manipulation.

KDSS commenced operations in 2000; the system was designed to monitor population change and to link that change to social, economic and environmental conditions in Kanchanaburi Province, which is located in the western part of Thailand. KDSS is composed of five strata: urban/semi-urban, rice, plantation, upland and mixed economy; each stratum contains 20 villages/census blocks. KDSS covers 80 villages and 20 census blocks in 13 districts of the province (Institute for Population and Social Research, 2001). Three sets of data-collection instruments, namely community, household and individual, are used in the annual enumeration of households and individuals. Spatial data, including the geographical location of each household, are also collected. Approximately 70,000 persons and 40,000 households have been enumerated and recorded in the system.

The initial data administration system was established for KDSS on the basis of "flat files". Such files were maintained separately for each year and for each unit of analysis, namely spatial, community, household and individual data. This database management system was developed on the basis of the long experience of the Institute for Population and Social Research (IPSR) with the collection and analysis of cross-sectional population-based data. The data were stored in SPSS (Statistical Package for the Social Sciences) format, and links among the different units of analysis and over different years were created through matching based on identification numbers.

As soon as the second year of data collection was completed numerous problems were identified in the database system. First, creating the linkages between the various data sets required an excessive input of human resources because of the complexity of the linking process. Second, problems relating to missing data became evident, especially in cases where individuals moved between study areas, and the identification number of individuals and households changed when they moved. Third, data retrieval was often inconsistent when matching data across years.

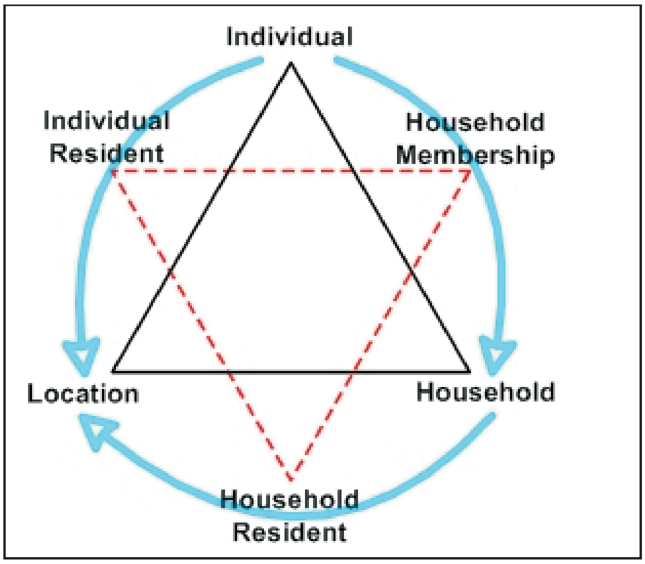
Moreover, as a member of the International Network for the Continuous Demographic Evaluation of Populations and Their Health in Developing Countries (INDEPTH), KDSS is interested in sharing data with other site members. The major barrier to sharing data is the different data structure of sites. The original data structure of the KDSS data set was a two-dimensional table containing 500-600 numeric variables. Often the data requested from INDEPTH required non-numeric data types, such as date and time data; however, KDSS had split the units day, month and year into three subfields and used a numeric data type for each field. The deficiencies of key fields, such as the start date when individuals appear in the study area and the end date when individuals terminated their stay there, resulted in time-consuming manipulations of the data.

Because of these difficulties, IPSR began the process of changing the database system in order to improve data administration and the reliability of the data. In this article, its experience in transforming the data management system to a relational database management system (RDBMS) is described so that the progress can serve as a lesson to other sites having similar needs.

Relational database management system

KDSS uses the INDEPTH standard approach in order to avoid data redundancy, diminish inconsistency and reduce the waste of resources and to produce a database that is compatible with collaborative database systems (Benzler and Clark, 2000; Benzler, Herbst and MacLeod, 2005). For these reasons, KDSS developed RDBMS, which is operated by SQL (structured English query language) (Lemsiriwongse, 2003). Owing to their concurrent multiple-user accessibility feature, relational databases satisfy many of the information needs of users and make it easier for them to design their own files (Gillenson, 1985). RDBMS is also compatible with spatial data operated on geographical information system software; thus, such data can be integrated into the database system.

Figure 1. Relationship among individuals, households and locations



The transformation to RDBMS was designed to separate the primary formatted data into constant or observation data types and classify related data sets systematically into tables. The related data could then be linked to other tables by created identification information. SQL is an efficient command language that is able to meet conditions and operate calculations simultaneously in order to retrieve data. Output of the SQL operation is compatible with data analysis software, such as SPSS; Stata, R; and SAS (Statistical Analysis Software).

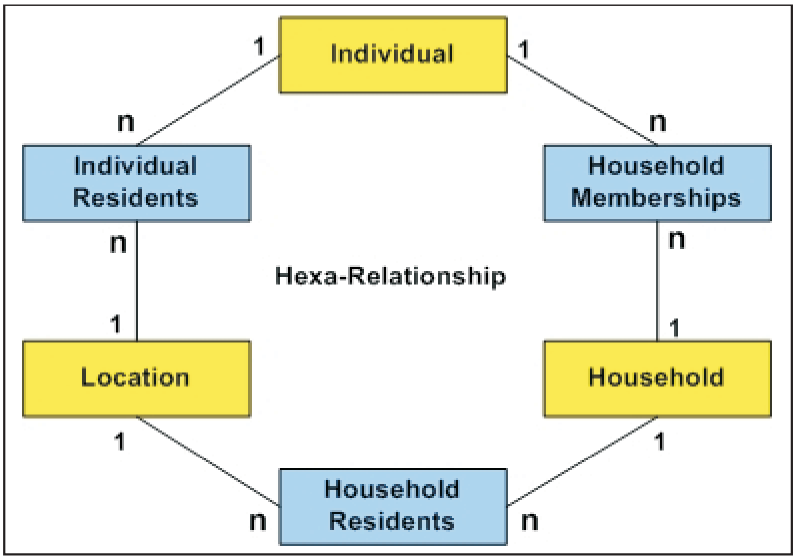
As a member of the INDEPTH Network (INDEPTH Network, 2002), the KDSS database was developed to be compatible with other demographic surveillance system sites on RDBMS (Benzler, Herbst and MacLeod, 2005) according to a demographic data structure called the entity-relationship model, which is depicted in figure 1. The structure includes the relationship among vital events that happen to individuals, households and locations.

A triangle with solid lines links the study objects: individual, household and location. The bottom-up triangle with dashed lines shows the relations or associations among three components: household membership, household resident and individual resident. The objects relate to one another. First, the arrow leading from the individual to the household signifies a relationship between the individual and the household, which is called household membership. Second, the arrow leading from the household to the location represents a relationship between the household and the location, which is called household resident. Third, the arrow leading from the individual to the location denotes a relationship between the individual and the location, which is called individual resident. Cardinalities are the possible numbers of relations; these are shown as a hexa-relationship in figure 2. These relationships simulate the relation of the population factors shown in figure 1. The cardinalities are modelled in RDBMS.

Cardinality (possible relations) can be classified into two types: one (1) and many (n). Figure 2 shows the cardinality between objects, which are described as follows:

- An individual refers to one person
- An individual may be a household member of either one or many households (n households)
- A household may have either one or many household members
- A household may move either once or many times to one or many locations
- A house/resident may contain either one or many households
- A house/resident may have either one or many persons living together
- An individual may or may not change locations

Figure 2. Hexa-relationship



Note: 1 = one and n = many.

The three related objects, individual, household and location, reveal relationships leading to the related vital event database design as well as implementation in the database system. The database structure design covers six necessary data sets shown in the hexa-relationship leading to the theoretical demographic database management system. Moreover, this design allows for future database expansion. The identification connectors, called foreign keys, are recreated, thus providing the means to join the tables. Consequently, the related data can be retrieved through the relation of tables.

Transformation to the new system

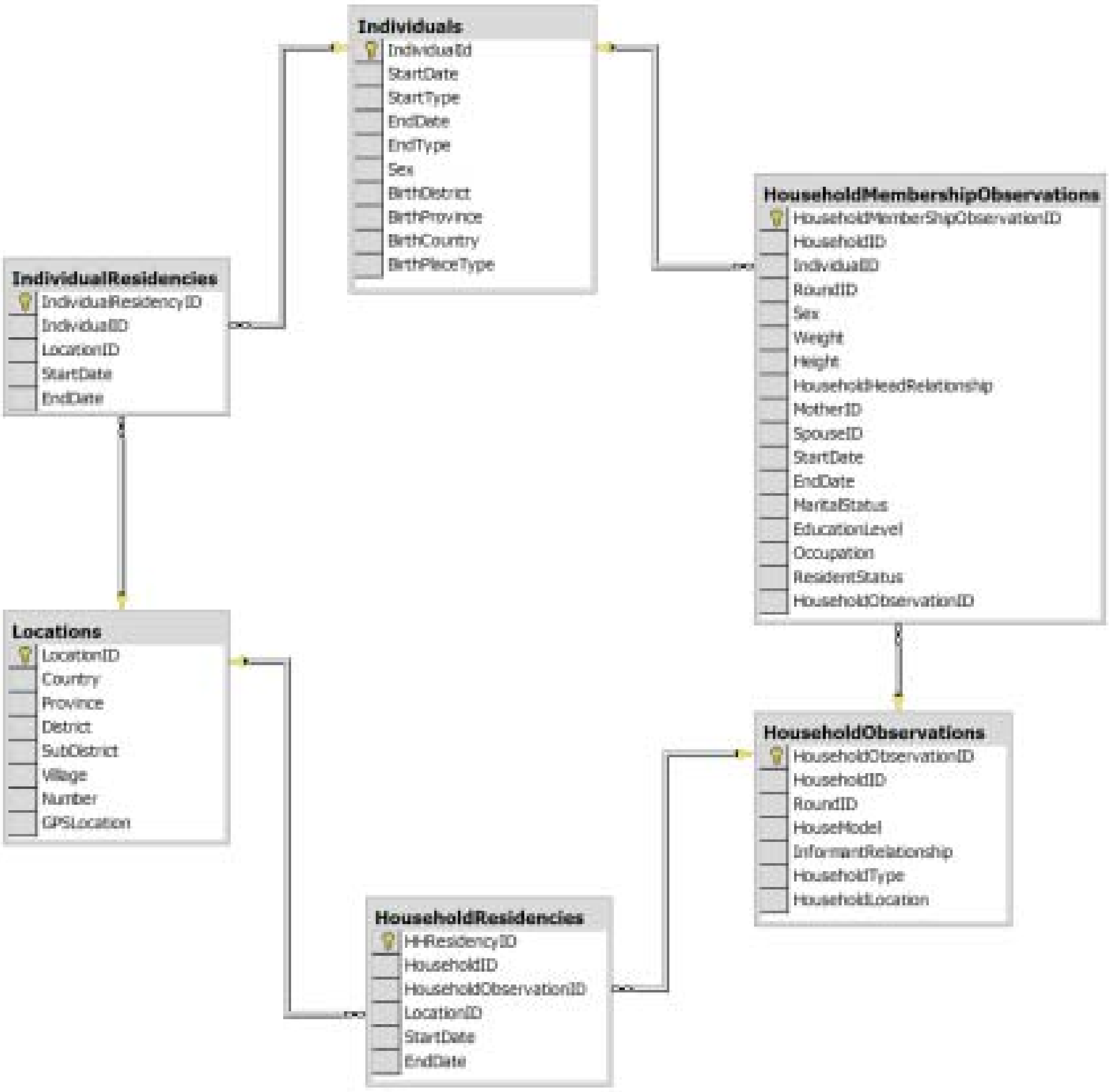
The transformation of the KDSS data management system from its original structure to the longitudinal scheme was costly and time-consuming. Table 1 shows the process and the human resources that were required. To accomplish the KDSS transformation to the longitudinal scheme, five steps were undertaken. First, a data investigation and feasibility study on the original data format and an evaluation of data diversity and data volume found that the new system could be operated in conjunction with the original data system. Since the KDSS data set is very specific and requires data in formats that can be used to analyse demographic change, the initial review took six months to determine the methodology for transforming the database. Meetings with the site leader and database staff were undertaken on a regular basis in order to understand researcher needs and potential constraints. Second, because the independently captured data from previous survey rounds needed editing in order to improve data quality and avoid data inconsistency, SQL programming and data correction on the original data set were required to ensure that the cleaned data were input during the transformation process. Third, the database design needed specialized skills that KDSS did not have; therefore, INDEPTH arranged for the Africa Centre for Health and Population Studies, University of KwaZulu-Natal in South Africa to provide technical assistance at the early stages of designing and reviewing the database design. Fourth, when people move to other study areas their individual identification consequently changes; therefore, a new, consistent and confidential identification link was created for use in RDBMS. This step consumed the most time of all, as it was necessary to link individuals over five rounds, which involved approximately 20,000 individuals for each survey round. Fifth, in transferring data to the KDSS relational database, it was necessary to develop an SQL program for automatically transferring the data with specific conditions. The total time for data transformation was 23 months; the process required the work of 20 persons, for a total of 460 person-months, to complete the task.

Table 1. Resources taken to transform the Kanchanaburi Demographic Surveillance System to the longitudinal scheme

Work	Resources		Persons in charge	Activities
	Time (months)	Personnel (persons)		
Data investigation and feasibility study	6	2	Information technology (IT) experts	Meeting at least once a month
Inconsistency checking and data editing	3	7 2	Data clerks IT persons	SQL programming Data correction in the original data set
Database design	1	2	IT persons	Meeting with and assistance from INDEPTH
Linking individuals between years (5-round survey)	9	2 3	IT persons Data clerks	SQL programming
Transfer data to the relational database management system	4	2	IT persons	SQL programming
Total	23	20		

The KDSS data management system consists of three activities: data capture, data retrieval and data updating. As previously discussed, all data in the original database were in SPSS format structured as two dimensions: rows and columns. There were approximately 40,000 records/rows for individuals and 12,000 records/rows for households in each round. Columns contained about 500-600 variables. For the new database, data capture transferred all the data to the new platform employed on RDBMS, which needed a relational data structure, that is, an entity-relationship model.

Figure 3. Conceptual entity-relationship model



Note: The nomenclature in the Kanchanaburi Demographic Surveillance System uses English characters with the first letter of each word capitalized; it avoids special characters and spaces for convenience when querying data by field names.

Individual and household data are included in the conceptual entity-relationship model because they are the data most often used. Nonetheless, community data can also be merged into the KDSS database management system by using a unique identification number on "ID", in this case, the village ID. One of the policies of the project emphasized data confidentiality; confidential information, including name, surname and other means of identification, is not disclosed to any of the researchers. Although the spatial data are bound to the location ID, the geographical individual ID and household ID are masked. Therefore, researchers are able to use the spatial data attached to the individual and household data, but the subjects concerned cannot be identified. To manipulate the new data system, invented IDs created for KDSS are also applied to related research projects. Figure 3 shows the conceptual entity-relationship model for KDSS, which is composed of the following tables:

- *Individuals*: individual information
- *HouseholdMembership Observations*: household membership information enumerated annually contains the relationship between the household's members
- *HouseholdObservations*: household physical characteristics that are enumerated annually
- *HouseholdResidence*: household residency information updated by the most recent enumeration
- *Locations*: global geographic positioning system (GPS) and administration references
- *IndividualResidence*: individual residency information updated by the most recent enumeration

All variables or fields are meaningful in English and each information table is linked by an ID. Figure 3 can be described as follows:

- IndividualID links *Individuals* to *HouseholdMembershipObservations* and *Individuals* to *IndividualResidence*
- HouseholdObservationID, created automatically by the computer system when new records are entered to the database, links *HouseholdMembershipObservations* to *HouseholdObservations*
- HouseholdID links *HouseholdObservations* to *HouseholdMembershipObservations* and *HouseholdObservations* to *HouseholdResidence*
- LocationID links *Location* to *IndividualResidence*

The baseline information, for example the constant individual information contained under *individuals*, is enumerated annually, verified and then recorded once in the database in order to avoid redundancy. Other changeable information is kept as historic information. Each record can be identified by the automatically generated episode ID and date.

The data structure and relationships shown in figure 3 form a basic prototype. The data administrator may design data relationships that are appropriate to meet the requirements of each researcher. An example of the data requests might demonstrate how well users can manipulate the data in the KDSS relational database management system if a data request, adapted from an INDEPTH request, asks for three data tables. First is an individual table containing information on individual ID, sex, marital status, date of birth, start date and end date of the individual's appearance in the study area during the period 2000-2004. Second is in-migration 2000-2004, containing information on individual ID, in-migration date and in-migration duration and on whether or not return migration occurred. Third is out-migration 2000-2004, containing information on individual ID, out-migration date and out-migration duration and on whether or not internal migration occurred. Based on the new system, the data requested can be classified into two groups of variables: constant and observation. The constant variables are individual ID, sex, date of birth and start date. The observation variables are in-migration date, out-migration date, in-migration duration, out-migration duration, end date and information on whether return migration occurred and whether internal migration occurred.

In the original data format, constant variables could be retrieved directly from those fields, except for the start date because it had not been created. However, the retrieval of individuals appearing in the study area at different times might be problematic without duplication. Individuals appear at the first enumeration of the demographic surveillance system survey, while others appear in the ensuing enumeration by birth or in-migration or missed enumeration. The database structure in the new system is designed to store data for distinctive individuals in the *individuals* table, which means that that table contains no redundant data. Observation variables in the new data structure are updated and modified when vital events occur, for example in-migration, out-migration and death. Consequently, event records are updated based on the most recent enumeration. Since an individual might have more than one event record, for instance people might move many times, the database structure can function individually for those events. Out-migration is an example of an event for which each individual could have no record, or one or more records. The out-migration events are recorded individually and modified by the most recent observation. An out-migration event is terminated when the individual returns to his or her place of origin or dies. The duration of out-migration can be calculated by the function of the difference between the start date and the end date. Information on residency status, including return status, is collected and this is attached to every enumeration. Although those variables are collected separately in tables, they can be linked by a connector called the foreign key. This data manipulation can be done at once with logical conditions and the calculation of the SQL function. This operation responds to the data requested. Our experience has been that the relational database management system for KDSS is less time-consuming than the original database management system, although it does require the skill of an SQL programmer.

The most important lesson learned from the KDSS data management experience is that an identification system should be specified clearly for every unit of analysis and each unit should hold the same identification until the demographic surveillance system is terminated. Therefore, an ID should meet these following qualifications:

- Uniqueness: an ID refers to one individual or one household only
- Confidentiality: an ID contains no direct identification of individuals or households

The KDSS identification system was employed as follows.

Individual identification

(a) *IndividualID* is the ID which researchers use for linking individuals between each year. The ID is unique; it does not contain any meaning that can identify the person concerned. Each individual holds only one ID.

(b) *IndividualExternalID* is the ID that reflects the location meaning. It is used for field work and is not disclosed to the public.

Household identification

(a) *HouseholdID* is the ID which researchers use for linking households between years. The ID is unique and does not contain any meaning that can identify the location of the household concerned. Each household holds only one ID.

(b) *HouseholdExternalID* is the ID that reflects the location meaning. It is used in field work and is not for public access.

Functioning and performance of the system

The database management system can be divided into operations that include data retrieval, documentation and recovery, and data accessibility. Retrieving data from the database can be performed by SQL command. The output data format is compatible with any of the commonly used spreadsheet software. SQL is the standard query language and is similar to other English programming languages; it can be copied and revised according to changing conditions. This reduces the time that must be expended when similar procedures need to be undertaken. In addition, SQL has a feature for linking all designed information from one table to another. Its cascade feature enables users to update and delete all connected tables at once.

Variables, ranges of values, validation system, retrieving data, inserting and updating data can all be managed with the appropriate tools. Data collection and capture should be prepared at the beginning of the project. A mandatory feature of the entire system is flexibility, which is needed to provide accessibility for new users.

Documentation describing meta data or a data dictionary is necessary for researchers who are not familiar with a data set. Meta data or a data dictionary should comprise all information for data utilization, such as variable name description, data type, range of values, period of data collection and conditions. If some values are coded, a code book should be attached using standard codes, such as the *International Statistical Classification of Diseases and Related Health Problems* (10th revision) and the national, provincial/district and subdistrict codes (ICPSR, 2005).

In case a database system "crash", it is extremely important to have a backup and recovery system, and this should be run regularly in and out of house. In addition, a log file containing the updated data history is very useful for rolling back in case there is a database crash.

Data accessibility is based on a set of permissions and a level of authorization to a person for data manipulation in RDBMS. Names, surnames, individual IDs and the household IDs of respondents containing direct means of identification are confidential. Therefore, KDSS keeps these data separately from others and grants data access permission to the data administrator only. Researchers are able to use individual IDs and household IDs but these cannot identify persons or households. Table 2 shows the examples of authorization of levels.

As shown in table 2, only the data administrator is in the administrator group and able to read (retrieve) and write (insert/update) all data. The field manager can only retrieve all data but cannot insert or update any data. The data administrator may set up more groups, such as researcher groups, and assign a permission level to such groups for accessing specified data tables in RDBMS. New users may be added to the existing groups or newly created groups and assigned a specific level of authorization for tables which they want to use in their research. All users need to be approved by the KDSS project director and/or institute's director before being granted accessibility.

Table 2. Examples of users authorized in the RDBMS

User name	Group	Authorization	
		Read	Write
Data administrator	Administrator	All	All
Field manager	Staff	All	None
Rittirong	Researcher	Specified data table, except confidential data	None

With the transformed database system it is convenient to retrieve data by writing SQL commands which can be revised and re-used. The unique identifiers enable efficient search through all the tables in the database because users are able to set conditions and they do not need to link individual or household annual records for each data retrieval requested. Consequently, this system reduces the time that must be expended and the number of processes required for data retrieval. Data consistency can be verified automatically by specially designed computer programs that run during data capture. As a result, data quality is improved.

Moreover, KDSS operating on RDBMS is flexible and is able to add more records to tables; the systemis also sufficiently flexible for adding values to newly created data description tables. Since the variable and value systems are intended to append new values instead of changing them over various surveys, users do not get confused with the expansion of variables.

RDBMS is efficient for managing a large volume of related data. Although it has no feature for advanced statistical analysis, all data retrieved from RDBMS can be exported to statistical software packages.

Conclusion

Database management plays an important role in KDSS: it provides data from the longitudinal data set that can be analysed and it improves data quality. The operation and access of the initial database system used in KDSS was costly and time-consuming. Therefore, a new system based on a relational database management system was developed to overcome these disadvantages. RDBMS operates by using structured English query language, which is reliable and sufficiently flexible for operating a longitudinal database. In addition, the KDSS relational database was developed based on the INDEPTH model; therefore, it is compatible for sharing data among other sites in the Network.

To formulate an RDBMS, technical issues must be incorporated within the system: in particular, an identification system should be specified clearly for every unit of analysis. Each unit must hold the same identification until the demographic surveillance system is terminated. Although RDBMS has no advanced statistical analysis functions, it is powerful and able to manipulate the data into formats that are accessible to users. RDBMS is able to update data history, and back up and recover data. These features minimize data damage in case the system crashes.

The experience of IPSR in creating the RDBMS database for KDSS is useful for other research projects that are developing longitudinal database systems. Our experience indicates that it is essential when creating any longitudinal database to invest in the development of systems that maintain confidentiality, while affording the basis for the numerous data linkages that are required for longitudinal data analysis.

References

Benzler, J., K. Herbst and B. MacLeod (2005). *A Data Model of Demographic Surveillance Systems: INDEPTH-Network*, <www.indepth-network.org/publications/indepth_publications.htm, accessed on 5 August 2005.

Benzler, J. and S. Clark (2000). *Longitudinal Database Design,* INDEPTH Workshop, 1999 Annual General Meeting of INDEPTH, Johannesburg.

Gillenson, M.L. (1985). "Trends in data administration", *MIS Quarterly,* December, pp. 317-325.

ICPSR. (2005). *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle,* Ann Arbor, Michigan, United States, the Interuniversity Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan.

INDEPTH Network (2002). *Population and Health in Developing Countries, Volume I: Population, Health, and Survival at INDEPTH Sites,* Ottawa, International Development Research Centre.

Institute for Population and Social Research (2001). *Report of Baseline Survey Round 1 (2000)*, Salaya, Nakhon Pathom, Thailand, IPSR, Mahidol University.

Lemsiriwongse, O. (2003). *Database Design and Management,* Bangkok, SE-ED Education (in Thai).